311 flooding complaints

Spatial Statistics Final Project

Research questions

- What areas of NYC see an unusually high number of floods? Are these areas typically surrounded by other areas with lots of floods? What about areas with very few floods?
- Are there any outliers in NYC (ie. places with very little flooding in areas with a lot of flooding or vice versa)?

Methodology

Extract and examine 311 flooding complaint dataset *125,955 records*

Used 311 NYC API to pull all records assigned to the Department of Environmental Protection with a descriptor variable that contains the word ‘FLOOD’ and a created_date variable that falls between 2016-2021. Some records are missing location data or are duplicates of other complaints so I had to drop them.

https://github.com/mghersher/flooding_hotspot_analysis/blob/main/01_311%20Extract%20Flooding%20Data%20and%20merge%20census%20data.ipynb

311 flooding complaints (2010-2021)

Why 2010-2021?

2010 is the earliest year that NYC provides data for. I chose to use all of it to help smooth out noise caused by things such as:

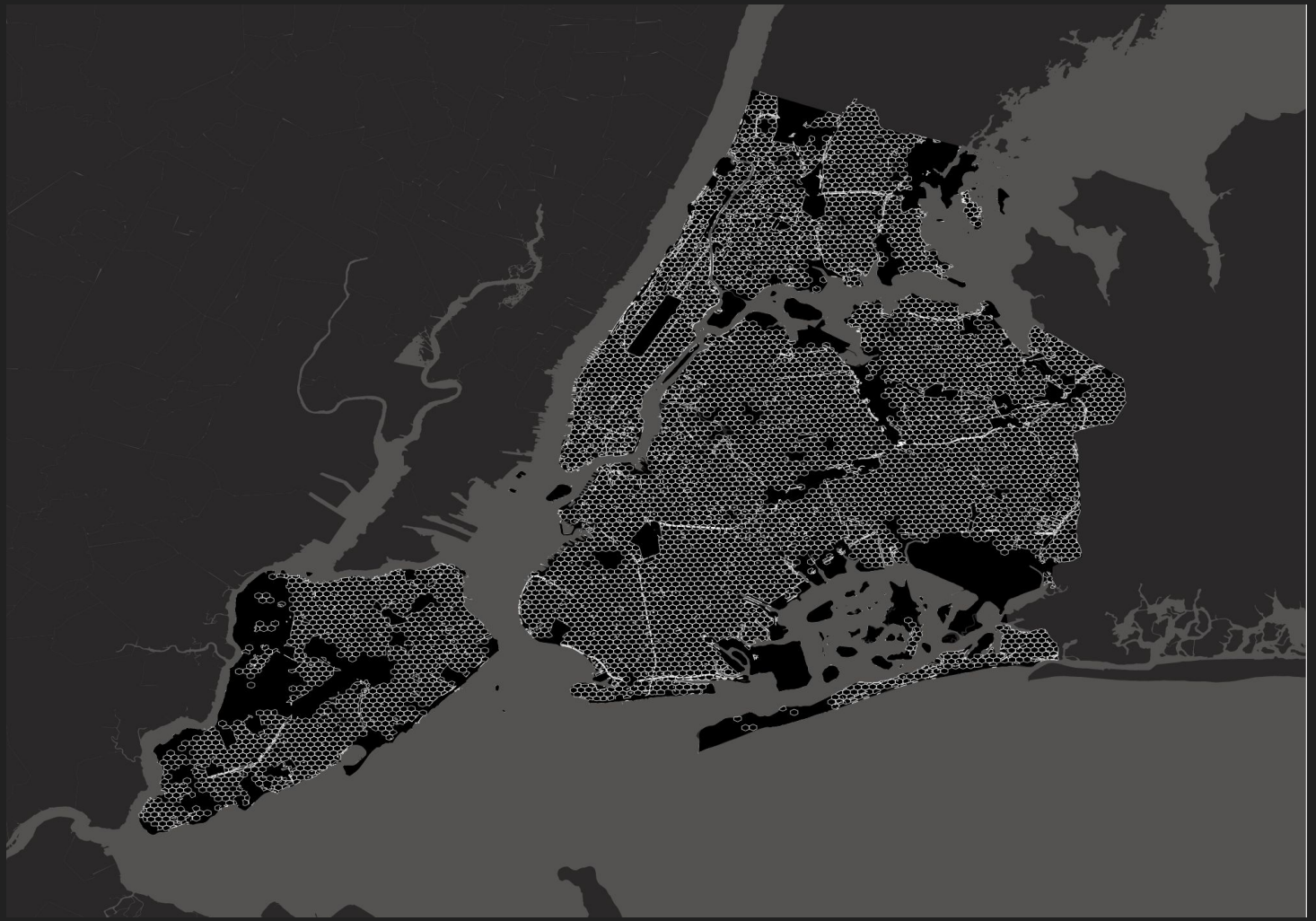
- Call operator recording error
- Caller reporting error
- Construction temporarily impacting drainage

Why is this well suited for spatial stats?

- The dataset is large! When you look at it even with a screen render to highlight areas with lots of overlapping complaints it's hard to discern hotspots and coldspots by eye. Different people would likely see different things.
- Looking at these questions statistically will let us discern reproducibly whether hotspots, coldspots, and outliers of 311 reported floods are due to meaningful variation in the processes underlying the data or random chance?

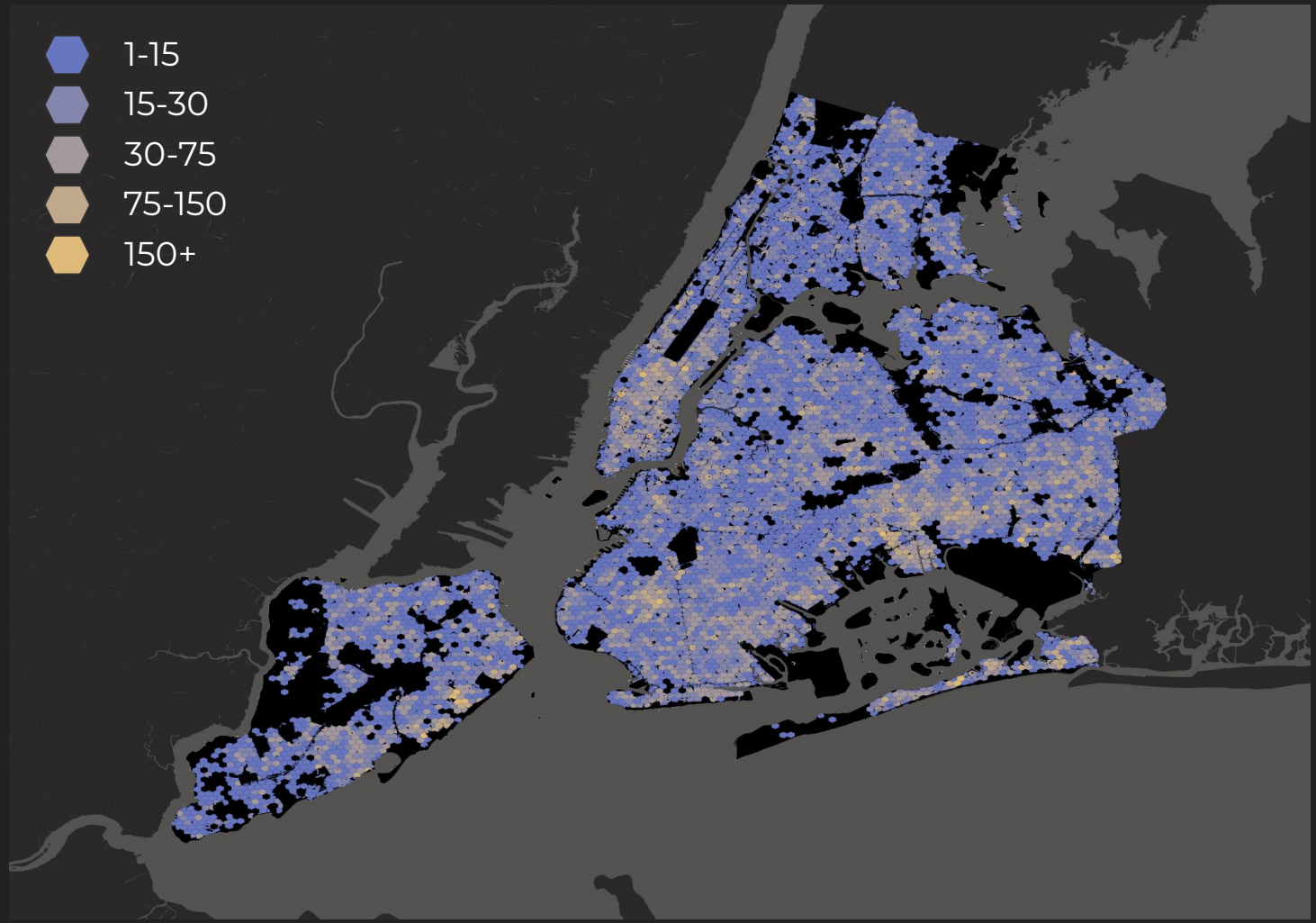
Hexagonal grid

Each grid cell is ~.04 sq miles



311 flooding complaints (2016-2021)

Count of flooding complaints in each grid cell



Normalizing flooding complaints by population

The 311 flooding complaint data is not a direct measure of where flooding has occurred across New York City. It is a measure of where ***reported*** floods have occurred. This means that floods that occur in more densely populated areas are more likely to be seen and reported. This means areas with lots of complaints in the 311 dataset might not be areas with a disproportionately large number of flooding incidents, but areas with lots of people.

One way we handled this was by dropping duplicate flood complaints, which deals with multiple people reporting the same flood. We still need to normalize by population though to handle the probability of someone seeing and calling about a flood.

What is needed to normalize the 311 counts?

The number of people who live in each hexagonal grid cell. This requires **areal interpolation**.

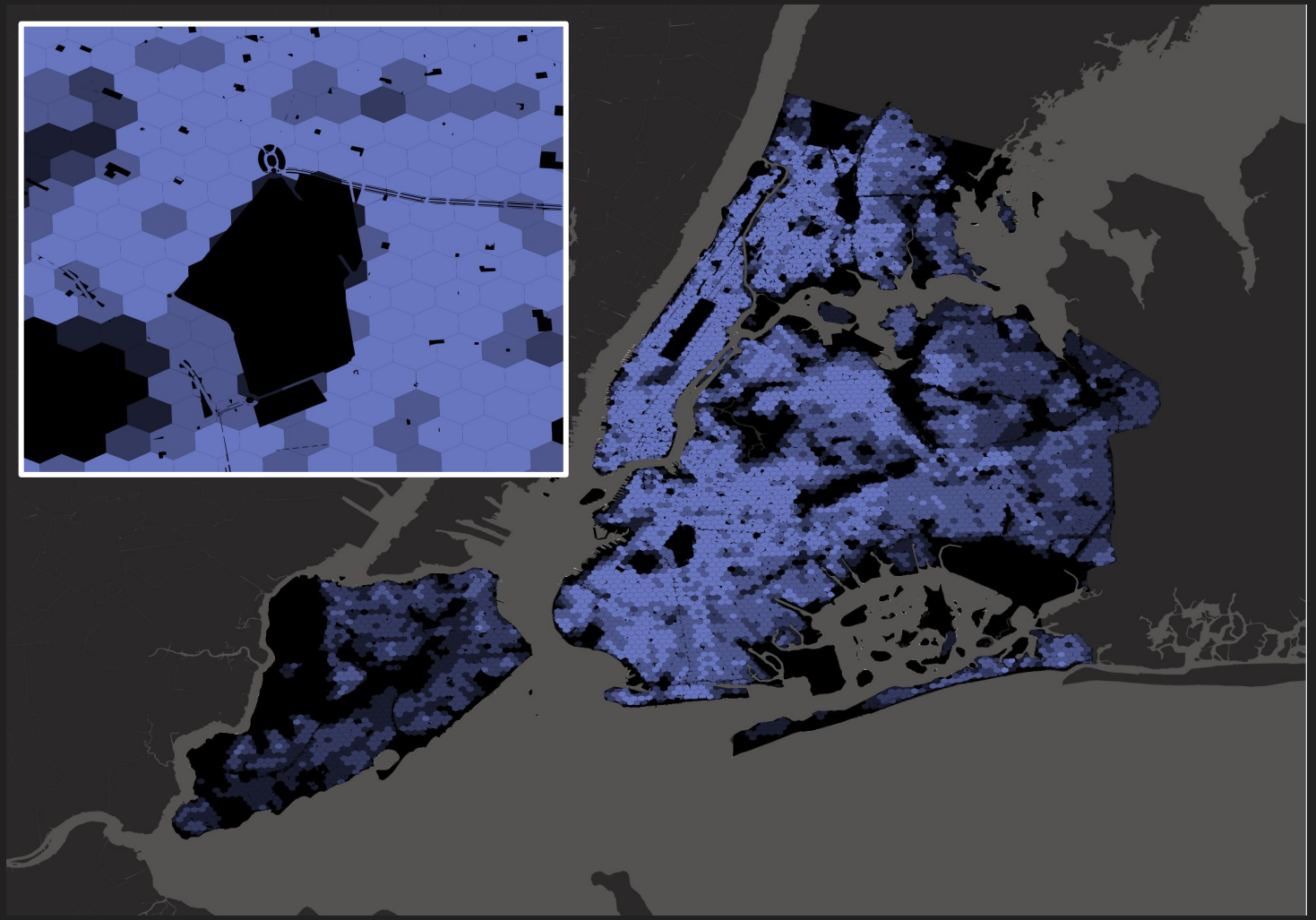
https://github.com/mghersher/flooding_hotspot_analysis/blob/main/02_Hex%20grid.ipynb

2020 Block-level Population

Merged 2020 redistricting census data with block shapefiles

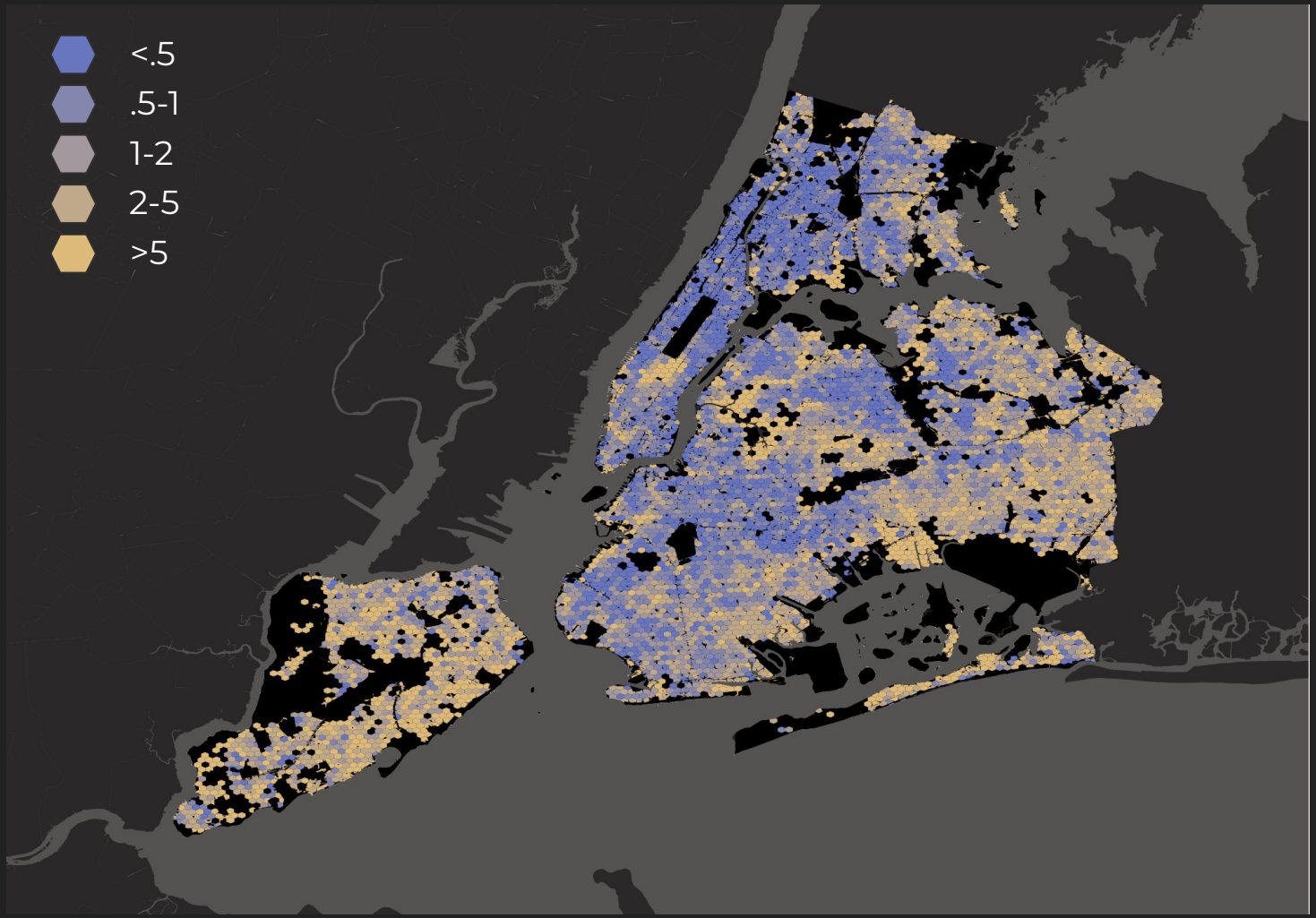
2020 Population

- <10K
- 10K-50K
- 50-100K
- 100-200K
- >200K



311 flooding complaints per 10,000 people

(2016-2021)



Note on data problems and outliers

Infinity problem

12% of grid cells have 0 people living in them, driving the number of complaints per 10,000 people to infinity. I examined these and most have very few complaints in them. I replaced the 0s with NAN to prevent this behavior, basically dropping these grid cells from the analysis.

The long, long right tail problem

Out of 8,211 hexagonal grid cells:

25% have fewer than .3 complaints per 10,000 people

50% have fewer than 1.1 complaints per 10,000 people

75% have fewer than 3 complaints per 10,000 people

90% have fewer than 7 complaints per 10,000 people

95% have fewer than 13 complaints per 10,000 people

99% have fewer than 150 complaints per 10,000 people

.5% have more than 1,000 complaints per 100,000 people and as many as 2 million

The long, long right tail problem cont.

These outliers with many, many complaints per 100,000 people are driven by hexagons with very few people in them (as estimated from the areal interpolation). The estimates are likely unreliable. So...

I needed to define a reasonable cutoff at which to discard hexagons (ie. if a hexagon has fewer than X people it is excluded from the analysis)

Defining the cutoff

12% of the hexagons have no people living in them.

Among the hexagons with nonzero populations, the median population is 76,000.

10% of hexagons have fewer than 250 people. **I chose to use this 10th percentile as the cutoff and drop all hexagons with fewer than 250 people from the analysis.**

Local spatial autocorrelation using Moran's I

Conducted both on the unnormalized and normalized 311 flooding complaint counts

https://github.com/mghersher/flooding_hotspot_analysis/blob/main/03_Hotspot_analysis.ipynb

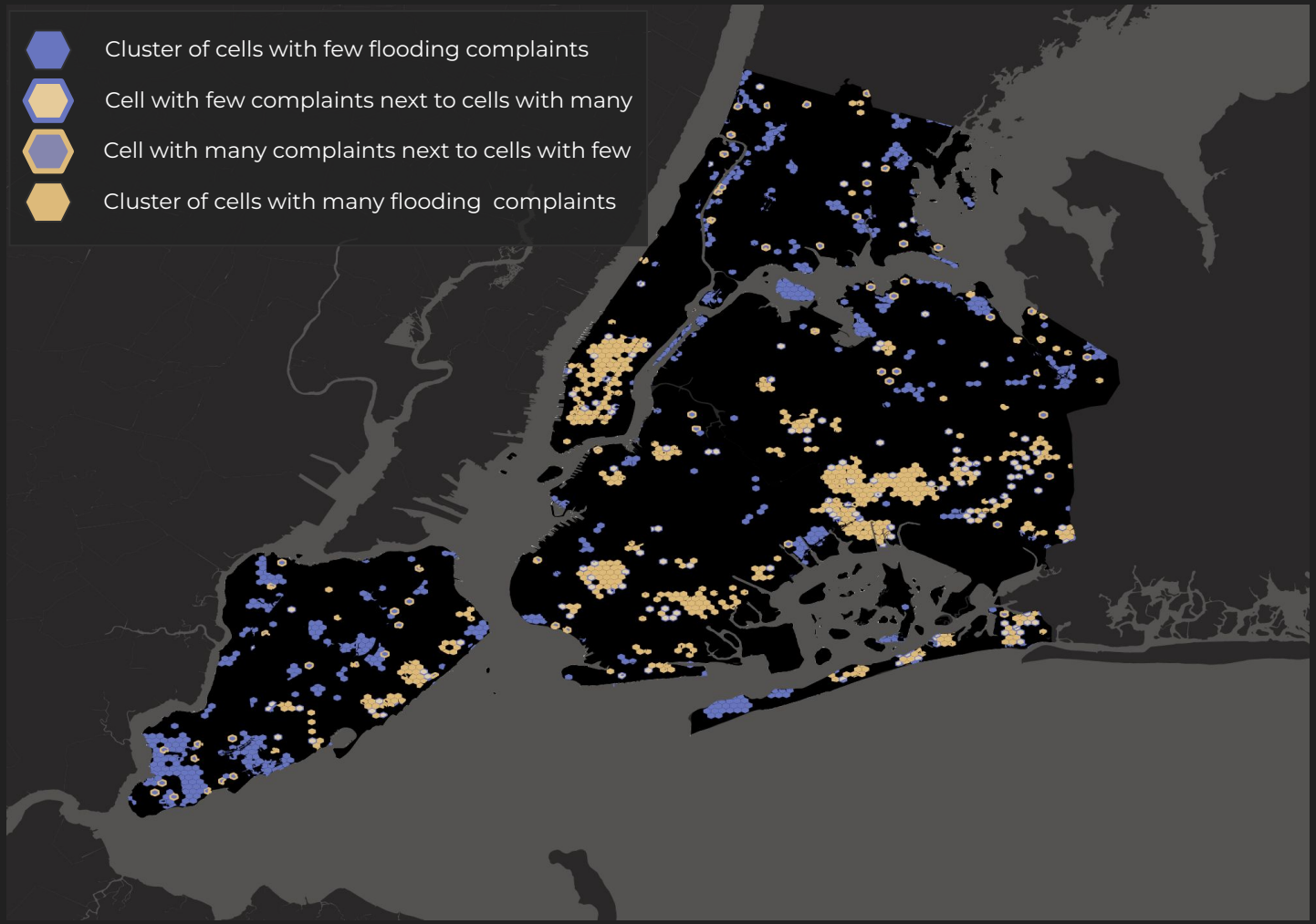
Used weights.KNN() and weights.attach_islands() to deal with hexagons that did not have adjacent neighbors

Results

Local autocorrelation

Moran's I

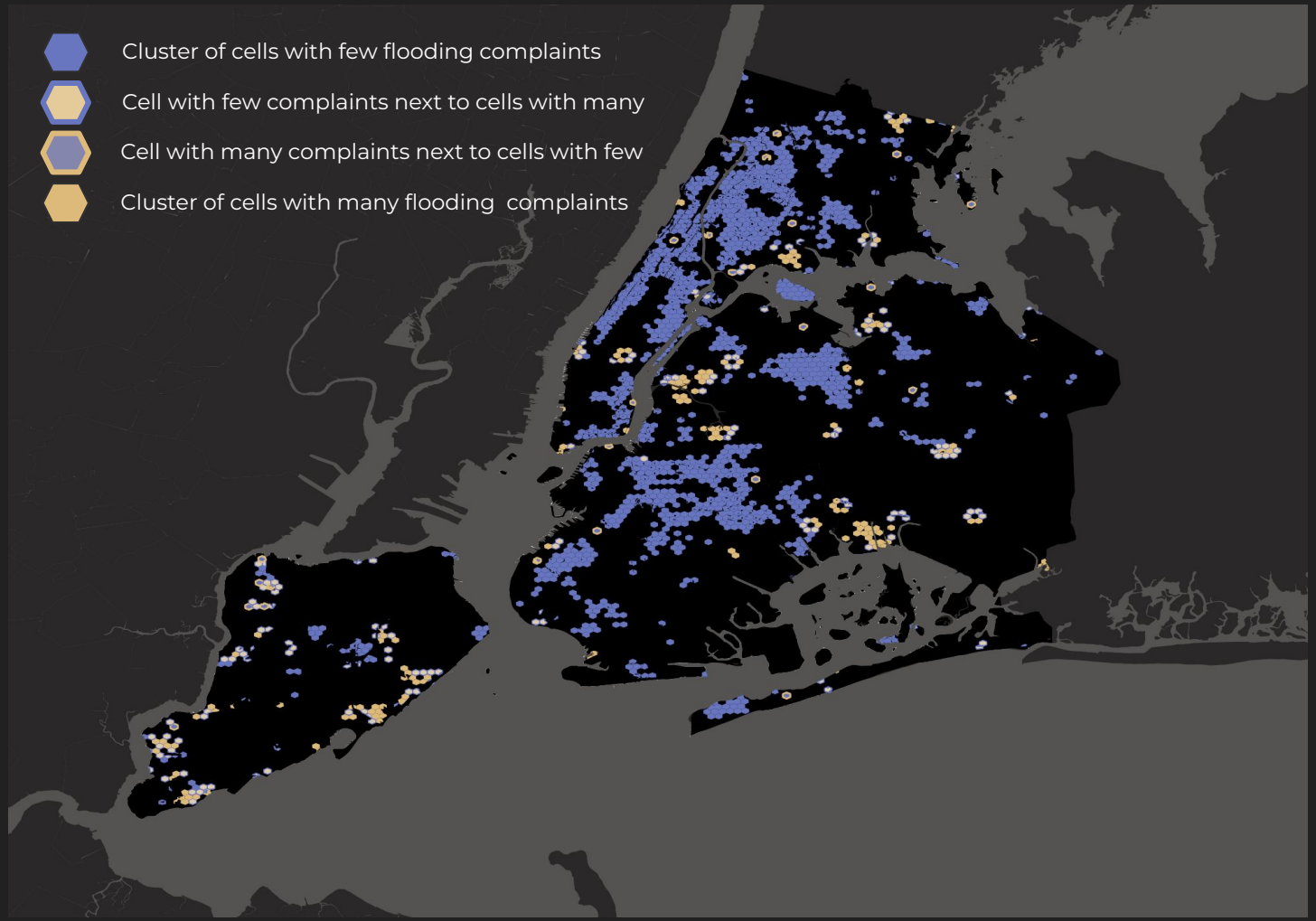
Executed on the unnormalized number of 311 complaints per hexagonal grid cell



Local autocorrelation

Moran's I

Executed on the normalized number of 311 complaints per 10,000 people in each hexagonal grid cell



Conclusions

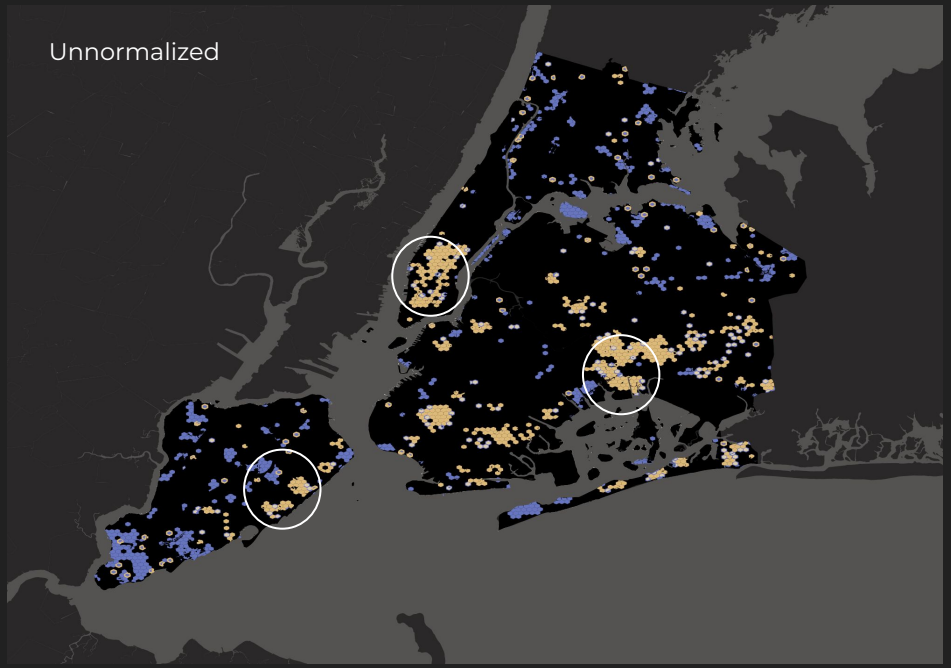
Most of the hotspots go away when you normalize the data. These are likely driven by high population, not a higher occurrence of flooding (ex. the large hotspot in lower Manhattan). Staten Island and Howard Beach in Queens have the most hotspots that remain after normalization.

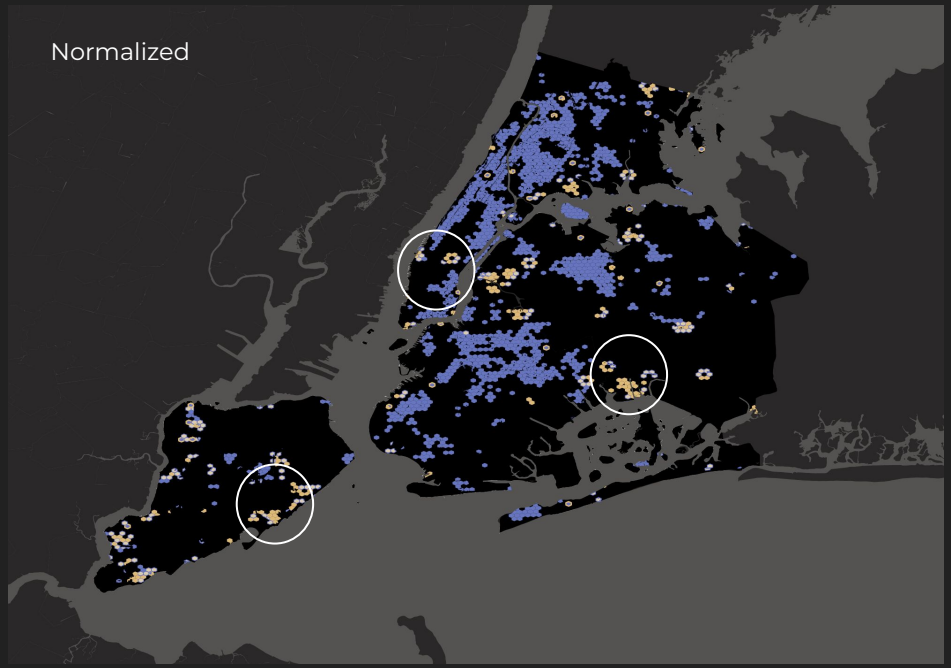
Cluster of cells with few flooding complaints

Cell with few complaints next to cells with many

Cell with many complaints next to cells with few

Cluster of cells with many flooding complaints





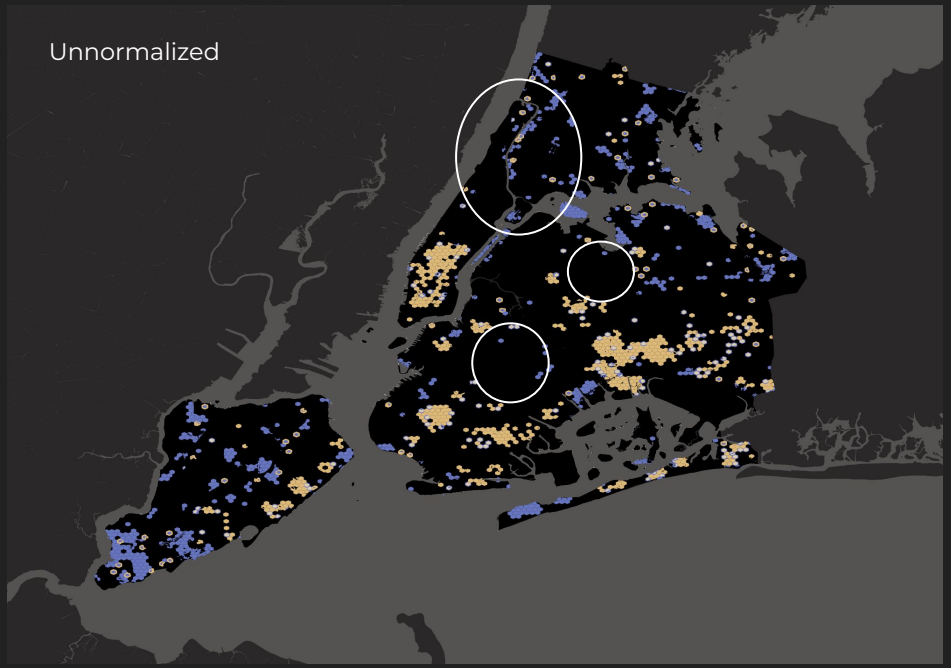
Very few coldspots disappear after normalization (with the exception of the southwest tip of Staten Island), but many not specifically significant areas before normalization become coldspots after normalization.

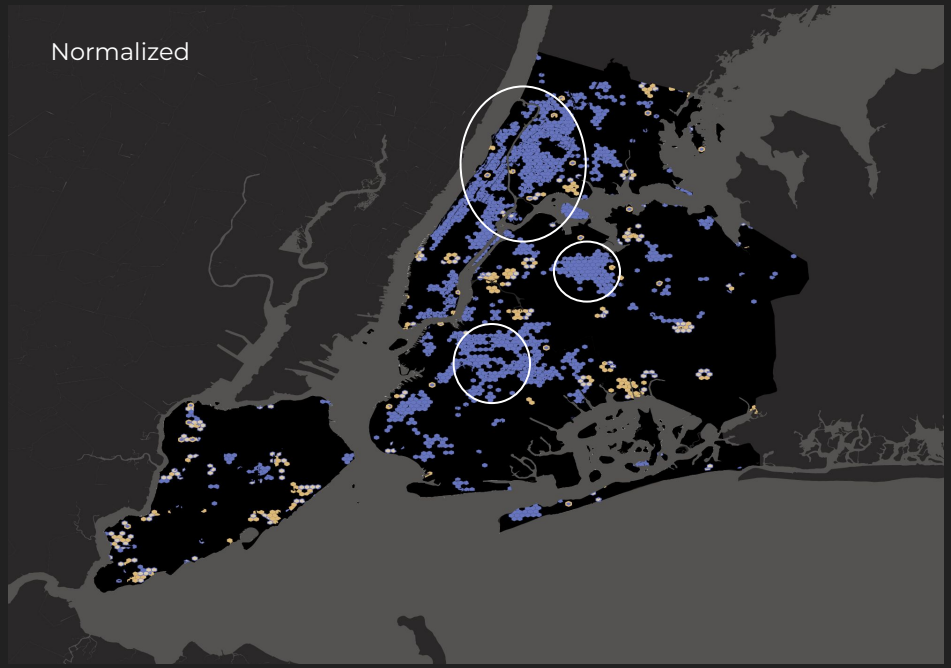
Cluster of cells with few flooding complaints

Cell with few complaints next to cells with many

Cell with many complaints next to cells with few

Cluster of cells with many flooding complaints





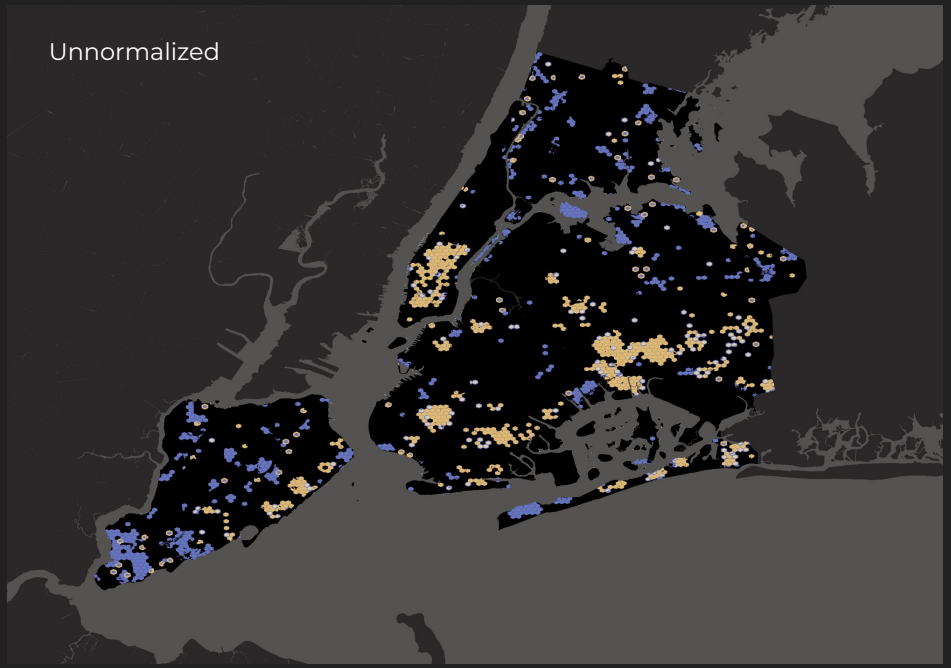
Almost all the diamonds and doughnuts shifted dramatically after normalization.

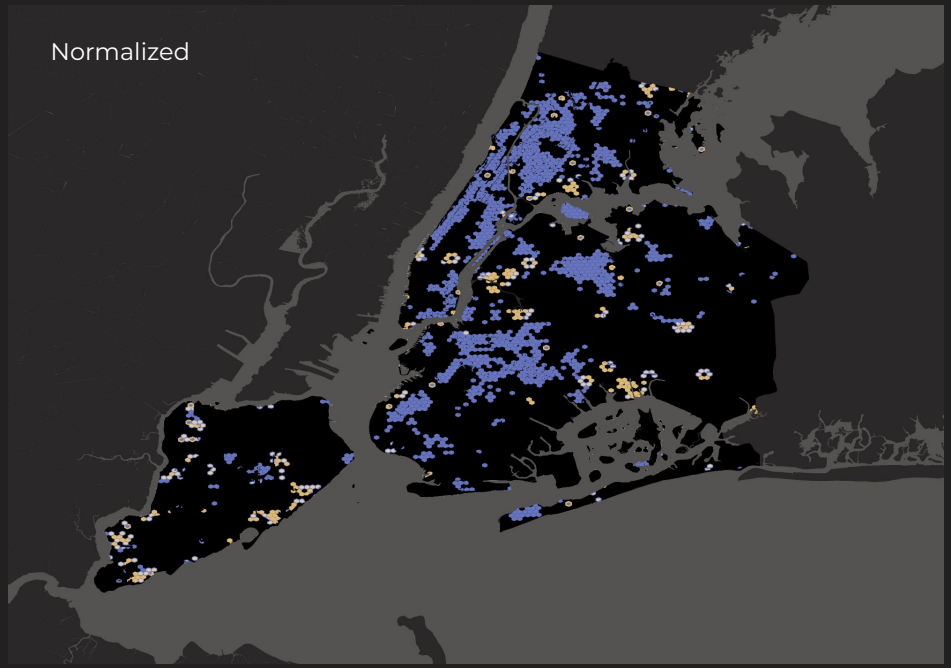
Cluster of cells with few flooding complaints

Cell with few complaints next to cells with many

Cell with many complaints next to cells with few

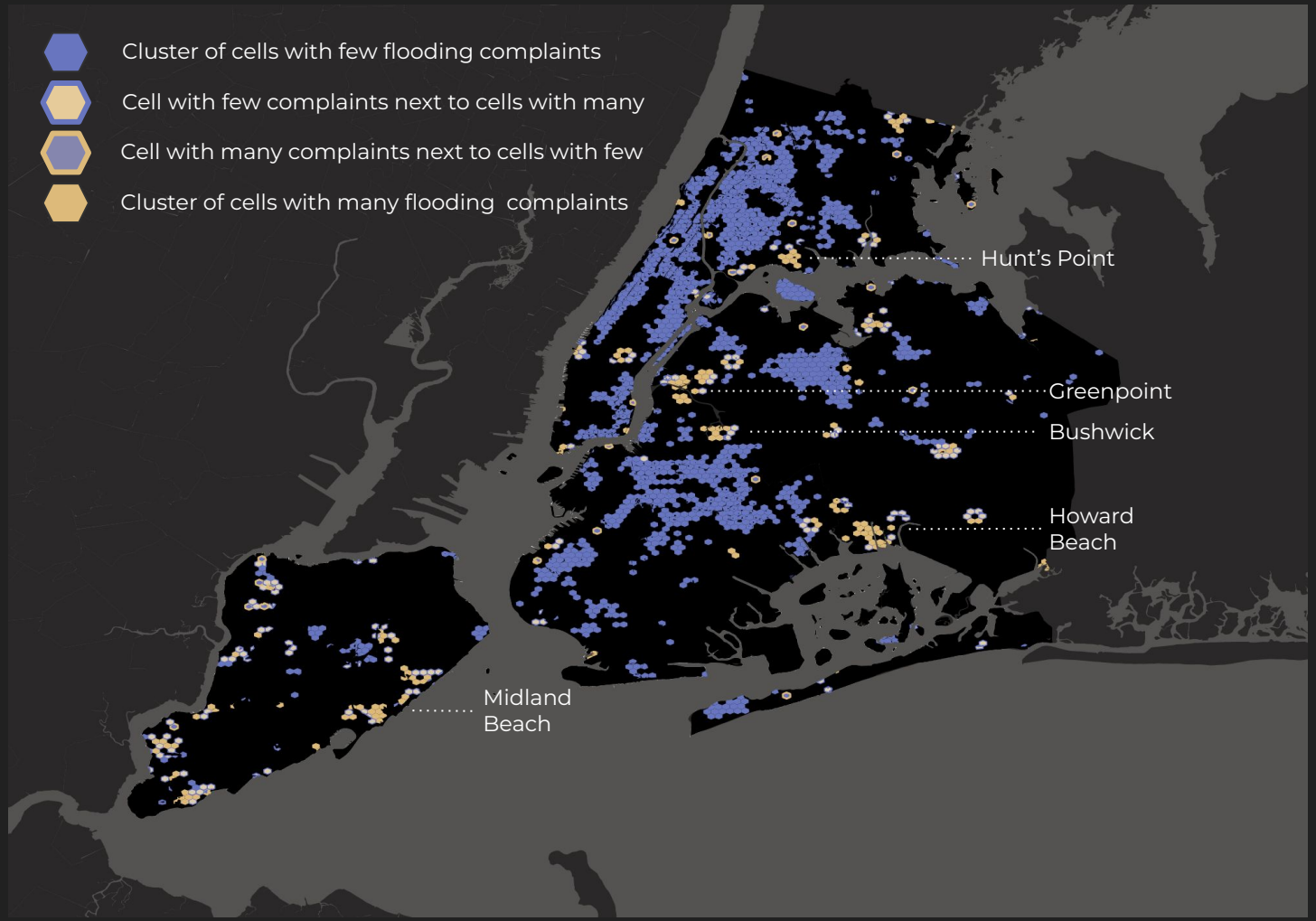
Cluster of cells with many flooding complaints



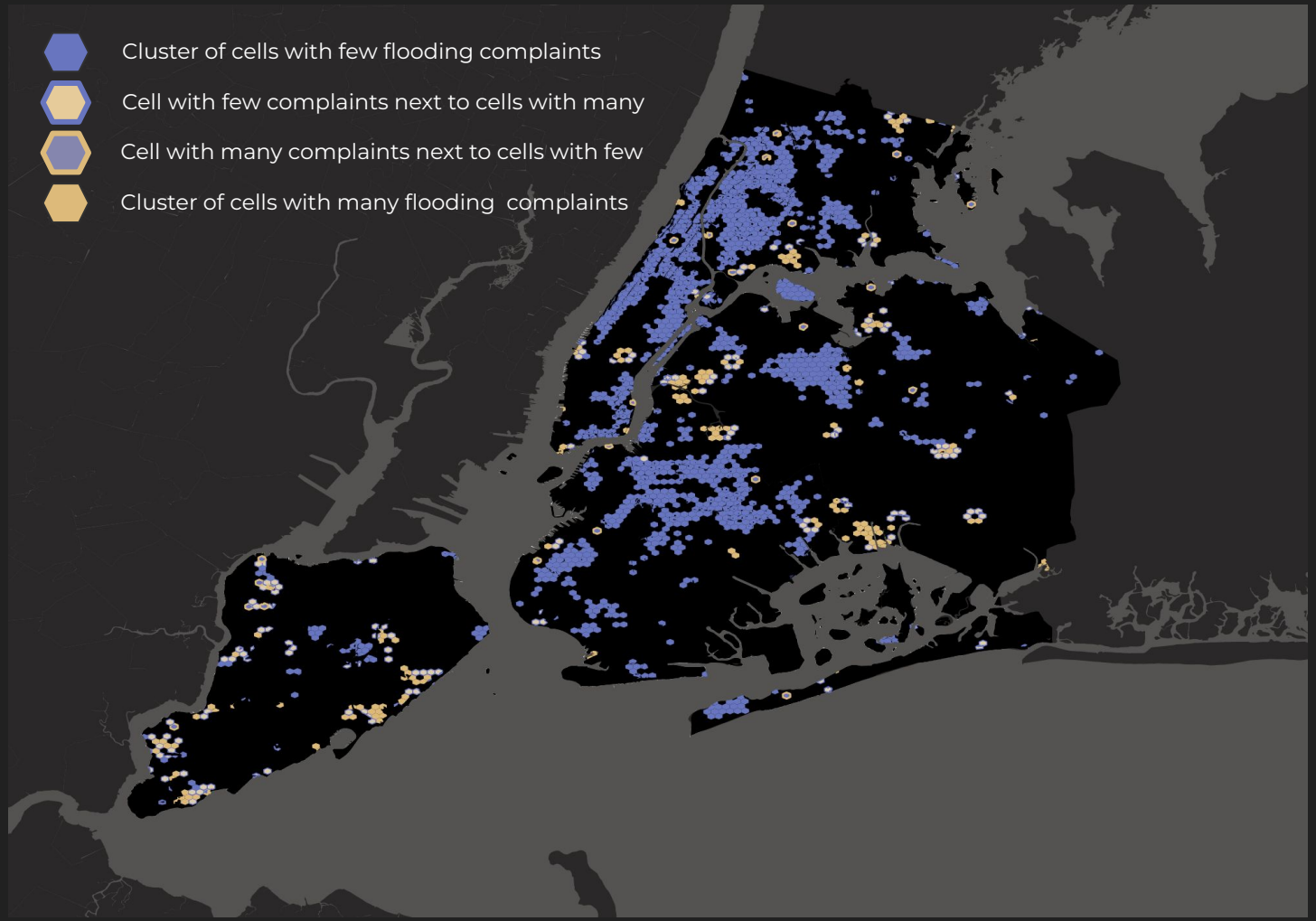


Some of the areas that have a high number of flood related calls per 10,000 people are unsurprisingly coastal (ex. Howard Beach, Greenpoint, Midland Beach, etc.).

Others though are not. For example, Bushwick has a cluster of cells with high a number of flooding complaints per 10K people. This is likely due to overflow from Newtown Creek draining to the East River

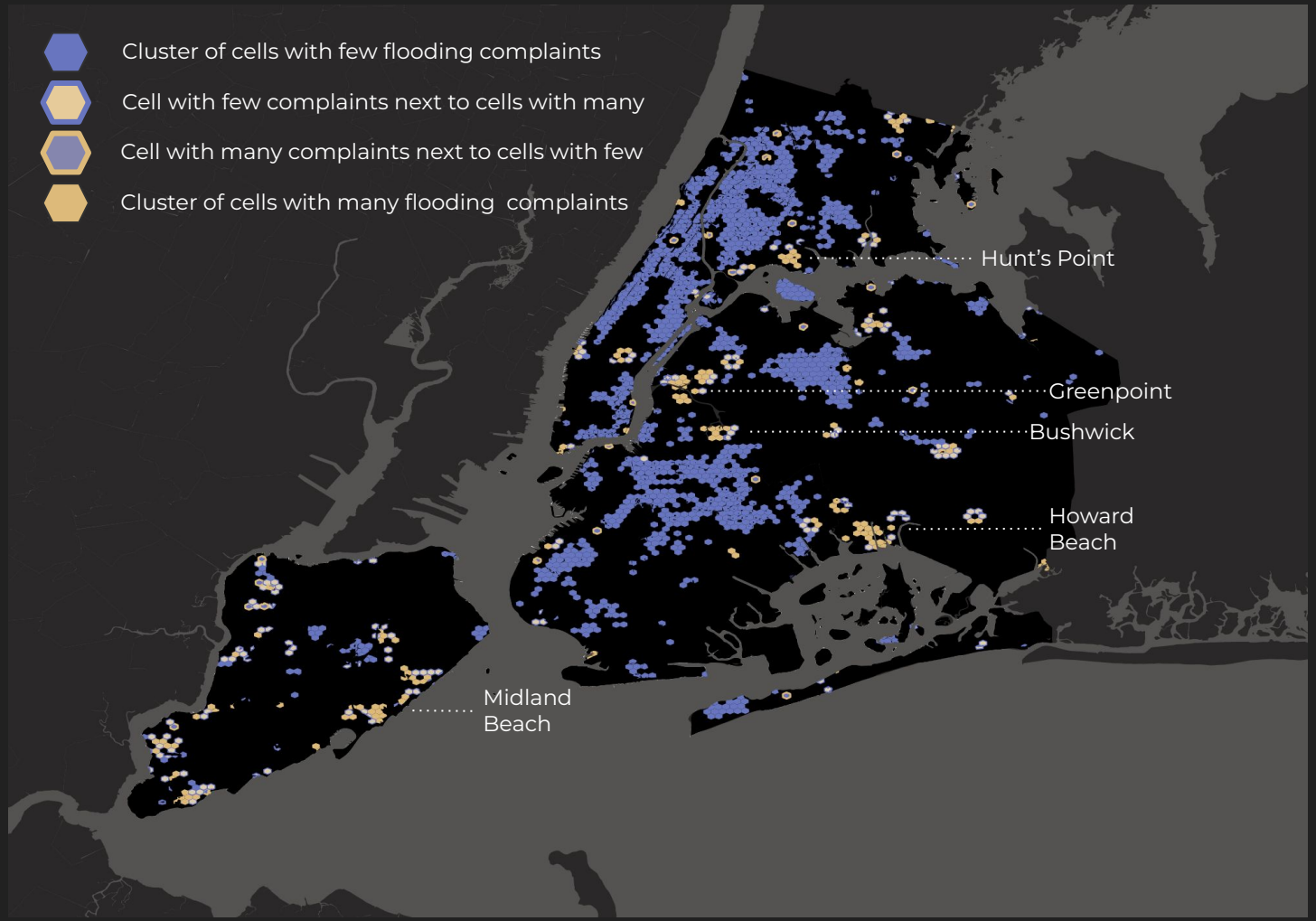


Many of the areas that look densely populated with 311 flooding complaints in the point map are actually coldspots. They - relatively speaking - have fewer 311 complaints logged per 10K people than New York City as a whole. Looking at a point map can be misleading!



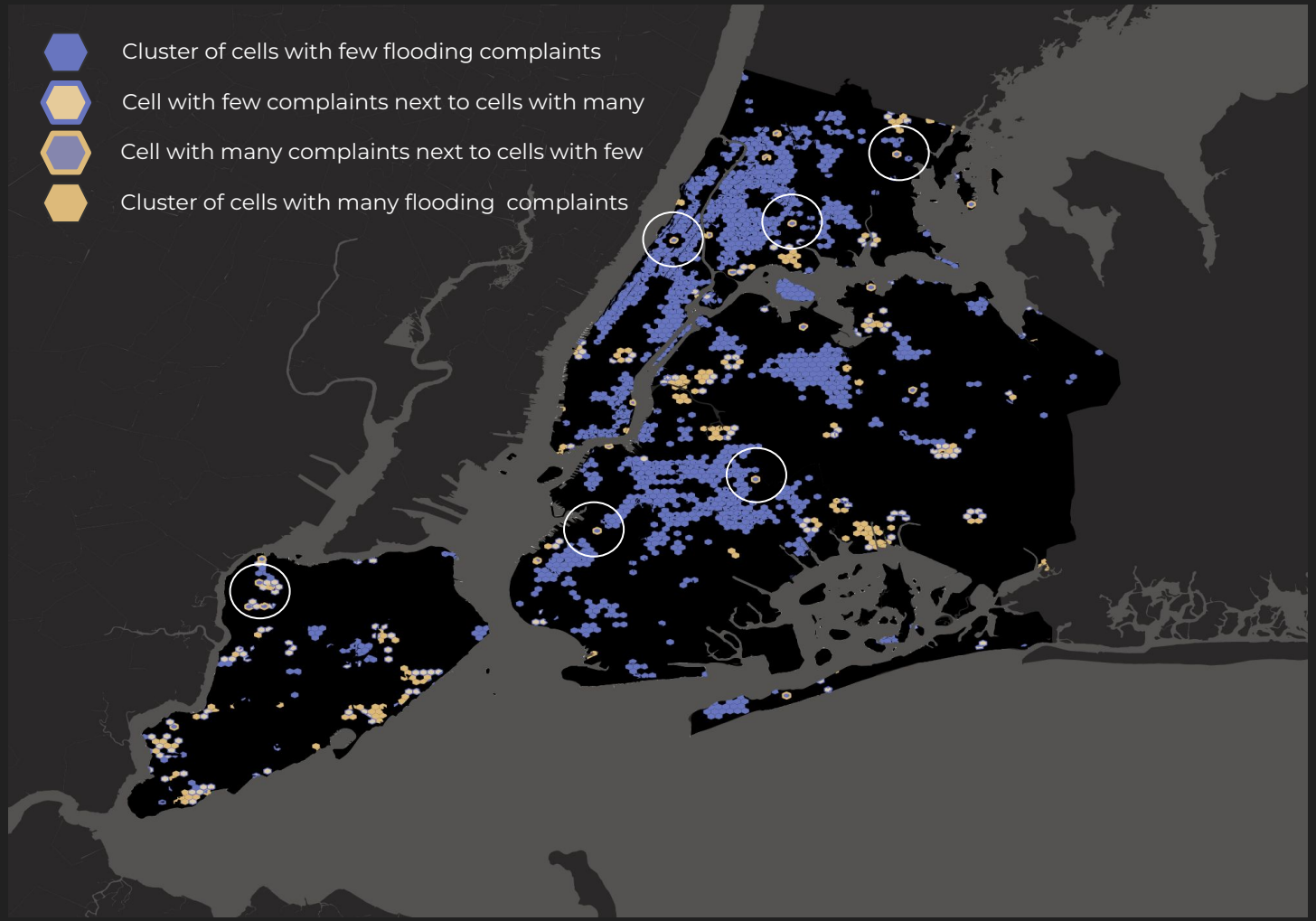
Many of the hotspots have doughnuts on their edges (ie. areas with relatively few 311 flooding complaints next to areas with a relatively high number of 311 flooding complaints.)

I would be curious to speak with a flooding expert to see if these areas have better drainage and whether they could be used as a model to improve flooding in proximate hotspot cells.



There aren't many statistically significant diamonds.

City planners could look to adjacent cells to see whether there are things that are different in these areas that could be improved in the diamonds to improve their drainage.



Future extensions

- Digging deeper into what causes the hotpots, coldspots, and outliers
- Could reaggregate other datasets into the hexagon to build features that might predict whether a cell is a hotspot, coldspot, or doughnut/diamond
- Could do a text analysis of the flooding description to distinguish types of floods and look at whether the clusters vary for different types of floods