Stat4601 Unsupervised Learning Map

What is unsupervised learning

- Where algorithms learn patterns and structures from unlabeled data, without explicit human guidance
- most common algorithms approaches : hierarchical clustering, k-means (that i will use them for this part)

How unsupervised learning helps in this project?

Since k-mean and hierarchical clustering are mainly use in this part, i will explain how each of them helps to discover the patterns.

K-mean

Pro:

- divided data into K predefined clusters
- requires specifiying K in advance
- captures borough-specific patterns, avoids forcing the same clusters on all boroughs

Why for Borough-level (local patterns) not global:

- Each borough has their unique market dynamics
- you can pick the number of clusters and capture local trends and easier to find the hidden trends and their own pricing models (cuz some of the price in that area might be consider as high in another area)
- Good and fast for small-to-moderate size

Hierarchical clustering

Pro:

- no need to predefine a K (cut the dendrogram at a desired level)

- builds a tree-like dendrogram of clusters

Why for NYC wide patterns (global) not local:

- no need to choose a fixed K upfront (since after combine all boroughs it will be a large dataset), Can reveal cross-borough patterns
- easier to see city-wide segments , like Brooklyn's fancy areas just like overall areas in Manhattan, or "In nyc, condos can divide into 3 types: luxury (manha + brookl), family-friendly (queen + bronx), budget (bronx or staten island)"

Workflow

- Apply PCA (not ICA because ICA mainly for image or whatever)
    - reduce dimensionality (avoid variables that are highly correlated)
    - speed up and stabilize algorithms like k-means
    - focus on most informative structure
    - Problems:
        - PCA can't run with any NA values, so in order to keep those NA value, i just make a df copy with complete dataset and work on them.
- Apply K-mean for each borough
- Apply hierarchical on nyc_data