Advanced Data Visualization & Data Cleaning

Using Seaborn & Feature Engineering with the Diamonds Dataset

Learning Objectives

By the end of this class, students will be able to:

- Create advanced Seaborn plots for urban datasets.
- Use pairplots for exploring neighborhood-level housing or zoning data.
- Clean and preprocess NYC PLUTO data.
- Engineer new features for urban analysis (e.g., density, volume).
- Use correlation heatmaps to understand patterns in urban form and property values.

Dataset Overview - NYC PLUTO

Source: NYC Department of City Planning (PLUTO)

- Parcel-level data for NYC real estate
- Contains building dimensions, land use, zoning, assessed values
- Useful for zoning analysis, development trends, and density studies

Variables we use:

- lotarea, bldgarea, numfloors, yearbuilt, zipcode, borough, landuse, assesstot

Violin + Strip Plot

- Explore the distribution of assessed property values by land use category
- **Use Case:** Identify how land use type (residential, commercial, industrial) affects assessed values
- **Why it matters:** Helps planners assess tax equity and land value patterns

Facet-Grid KDE Plot

Objective:

Visualize building height distributions across boroughs and zoning types

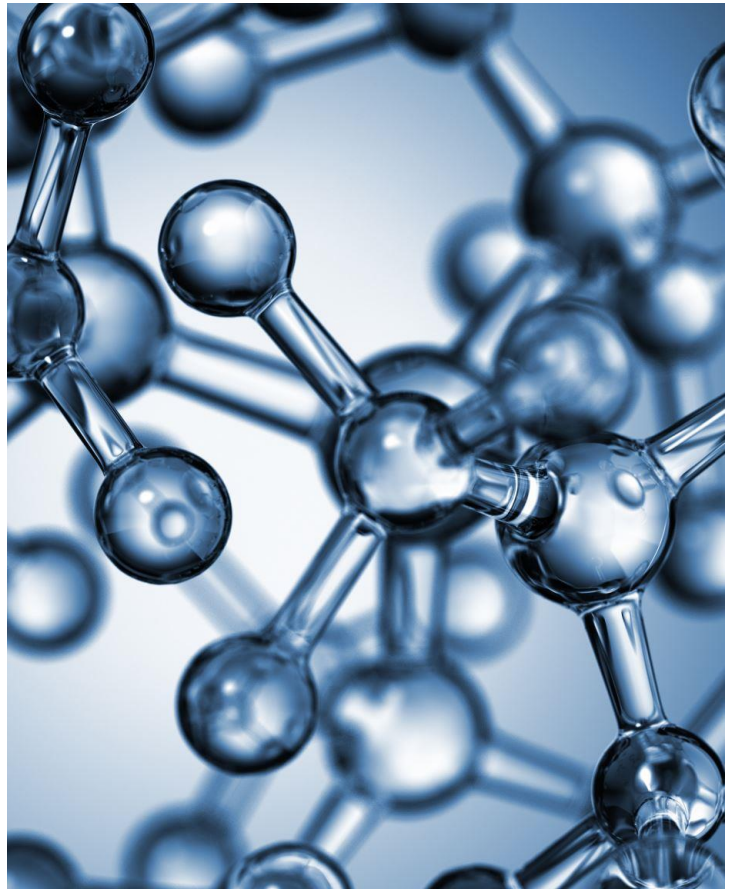
Use Case: Understand density trends and vertical development across NYC

Pairplot - Exploring Relationships

Objective:

Explore how lot area, building area, number of floors, and assessed value interact

Why it matters: Supports site selection and policy analysis based on built environment characteristics



Data Cleaning - Missing Values

Objective:
Handle missing values in key fields like *numfloors, yearbuilt, assesstot*
Strategy:
Drop extreme outliers
Fill missing with median or grouped means by borough

Feature Engineering - Creating Useful Features

Objective:
Create variables for density and urban morphology:
Floor Area Ratio (FAR): bldgarea / lotarea
Estimated Volume: lotarea * numfloors
Decade built: extracted from yearbuilt

One-Hot Encoding for Boroughs or Zones

Objective:

Convert borough, landuse into machine-readable format

Why: Useful for modeling and clustering analysis on zoning or tax distribution

Heatmap - Visualizing Correlation

Objective:

Identify strong relationships among built environment variables and assessed values

Use Case: Inform planning regulations, equity analysis, or site evaluation

Summary of Concepts

Seaborn plots: violin, KDE, pairplot, heatmap

Data cleaning: fill missing values, drop bad rows

Feature engineering: FAR, volume, landuse encoding

Discussion Prompt: How might these metrics influence decisions about rezoning or infrastructure investment?

References & Tools

References & Tools

- NYC PLUTO: https://www.nyc.gov/site/planning/data-maps/open-data.page
- Pandas: https://pandas.pydata.org
- Seaborn: https://seaborn.pydata.org