DS-GA 1007 PROGRAMMING FOR DATA SCIENCE COURSE PROJECT

Project Title: Analyzing NYC Property Sales for 2003-2022

Group "pip install Grade A"
(Chloe Zheng & Rodrigo Kreis de Paula)

November 2022

Today's Agenda

1 Introduction & Motivation

2 The Data

3 Methodology

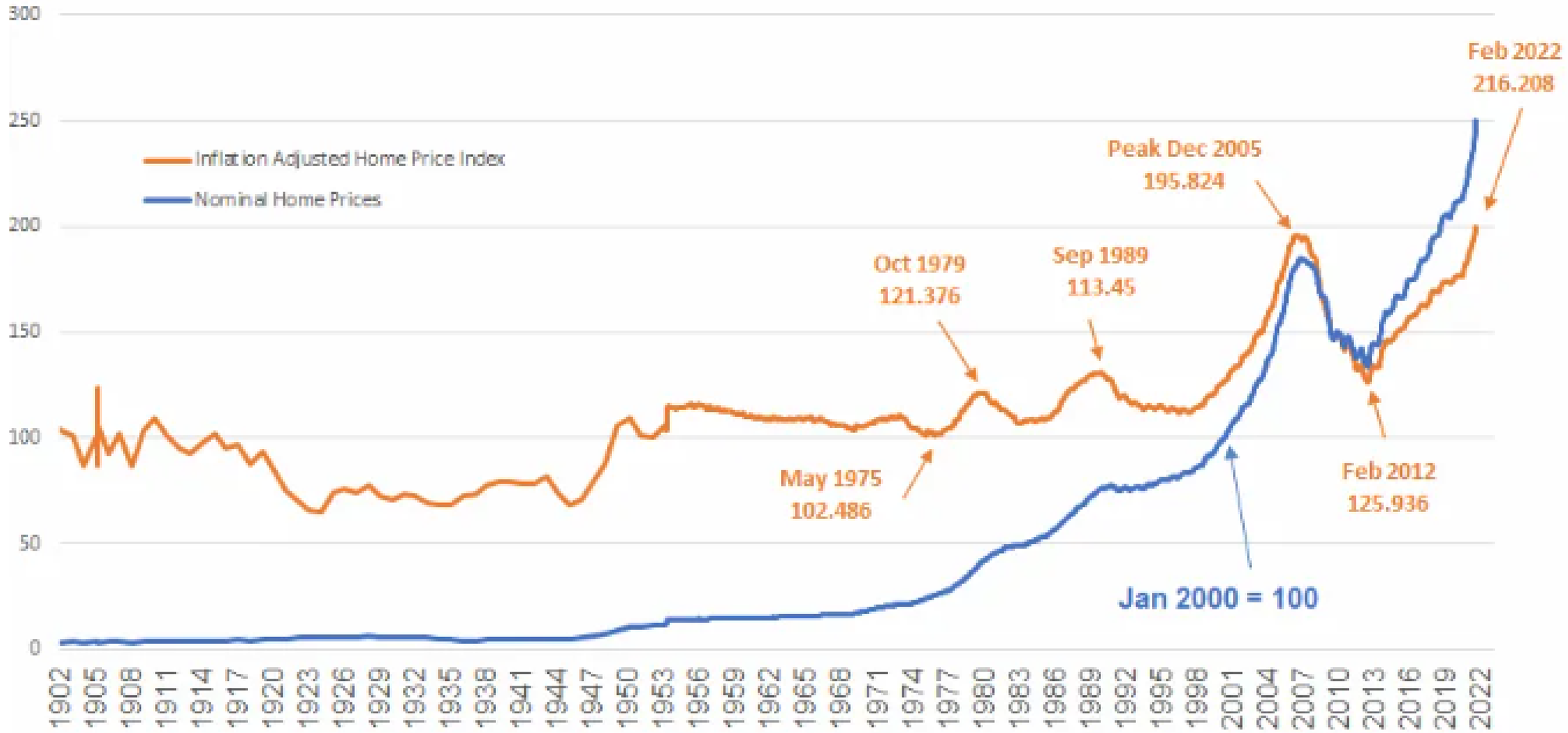
[Intro] "Property Sales" has so many implications

- business incentives
- research the housing crisis
- housing prices vs. inflation
  - US average rent rise 18% over the last five years
  - In 2020, 46% of American renters spent 30% or more of income on housing

- Disparities between privileged/unprivileged groups
  - 2021: 74% of White adults owned a home, compared with 43% of Black and 48% of Hispanic

- most of us (us included) want to buy our own property one day

[Intro] Housing Prices have been increasing in real terms



Values in USD thousands

Link to the graph here

[The Data] An "n by n" dataset

Dataset: NYC Property Sales for 2003-2022 (link)

A record of every building or unit (apartment, etc.) sold in NYC over the last **2 decades**.

21 features:

- **Location:** borough, neighborhood, block, lot, address, apartment #, ZIP code
- **Qualitative:** Building Class Category, Building Class at present and time of sale, Easement
- **Quantitative:** # of units (res/com/tot), land & gross square feet, year built, sale price
- **Temporal:** sale date
- **Tax:** tax class at present, tax class at time of sale

NYC 5 boroughs

20 independent variables
1 dependent variable (Sale Price)

1.83mm registers
1.28mm non-null "Sale Price" registers

[Methodology] Putting DS-GA 1007 in Practice

What we want to investigate:

- Most affordable locations in NYC
- Which features have higher/lower influence on prices
- Trends over the years

Methods:

- Data Gathering & Concatenation (multiple sparse files)
- Data Cleaning (null values)
- Feature Transformation/Engineering (including one-hot encoding, date/time)
- Distribution of independent and dependent variables
- Data Visualization (histograms)
- Features Correlation (heatmap)
- Group by Features (comparison of Boroughs / similar properties)
- Modifications in variables over time

Thanks!

Questions?