DS-GA 1007 PROGRAMMING FOR DATA SCIENCE COURSE PROJECT

Group "pip install Grade A"

by Chloe Zheng and Rodrigo Kreis de Paula

October 2022

Agenda

01 Introduction & Motivation

02 Data

03 Methodology

Introduction & Motivation

Topic: Real Estate

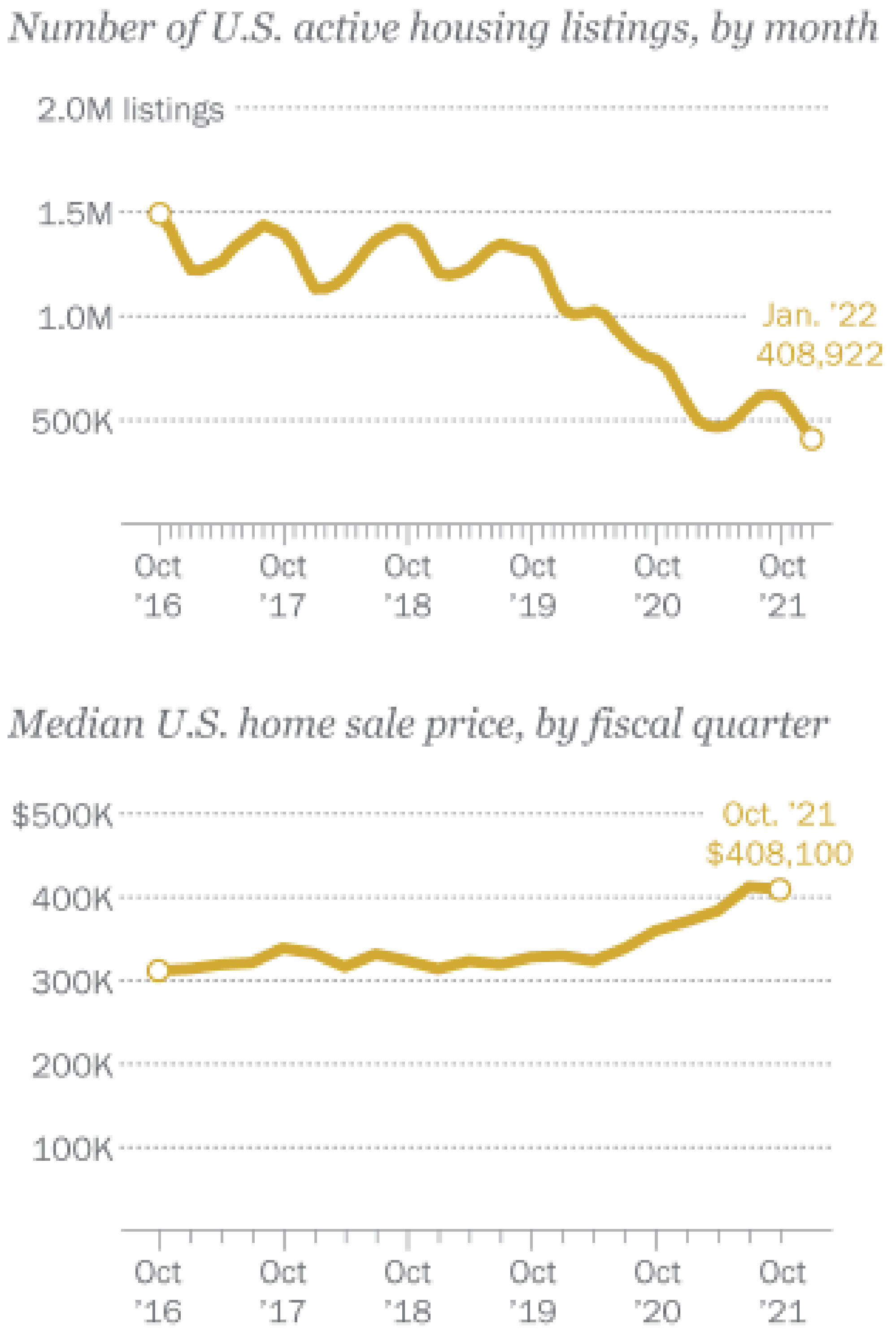
Why is Housing Data Important?

- Business incentives
- Research housing crisis
- Investigate rising prices and 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
- Investigate disparities between privileged/unprivileged groups:
  - 2021, 74% of White adults owned a home, compared with 43% of Black and 48% of Hispanic.

What we want to investigate:

- Understand most affordable locations in NYC
- Understand what features of a unit influence higher or lower prices
- Investigate trends from over the last decades
- Real application: most of us (us included) want to buy our own property one day 😥

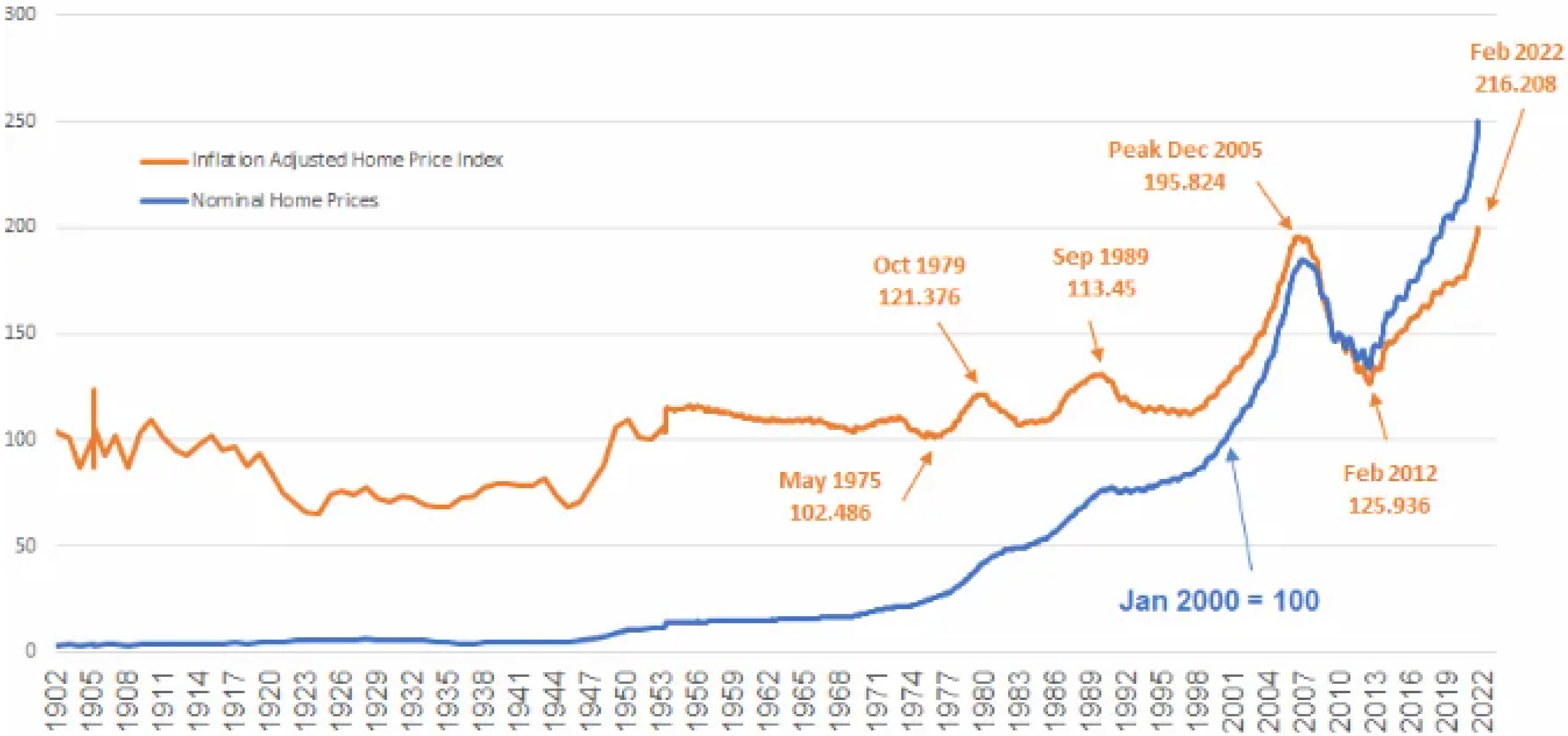
Home inventory is down, home prices are up



Source: Federal Reserve Bank of St. Louis.

PEW RESEARCH CENTER

Introduction & Motivation



Values in USD thousands

Link to the graph here

Data

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

The dataset is a record of every building or building unit (apartment, etc.) sold in the NYC property market over the last **2 decades**.

- **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
- **Temporal:** sale date
- **Tax:** tax class at present

NYC 5 boroughs

19 independent variables
1 dependent variable (Sale Price)

k (?) registers
70k non-null Sales Price

Methodology

- Data Gathering, Concatenation (multiple sparse files)
- Data Cleaning (null values)
- Feature Transformation/Engineering (including one-hot encoding, date/time)
- Distribution of values per feature (sales price)
- Data Visualization
- Features Correlation
- Group by Features (comparison of Boroughs / similar properties)
- Assessment of variations over time
- Extra - out of the course's scope
    - Displaying the information geographically (GIS mapping)
    - Sale Price prediction (regression)

References

- https://markets.businessinsider.com/news/stocks/new-york-city-housing-market-rent-facts-2019-6-1028269134
- https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page

Thanks!

Questions?