Visualizing NYC 2017 House Sales Data

Zhengqing (James) Chen

NYCDSA Cohort 12

01/31/2018

Outline

1. Project Objectives
2. Data (Source, Pre-Processing, Transform)
3. Shiny App Demonstration
4. Additional Observations

Project Objectives

- To visualize NYC house sales pattern, *borough-wise*

  In particular:

  - Transaction volumes (number of houses sold)
  - House prices

- To find additional insights from the above visualization
- To practice using R, Shiny, Leaflet

Part II. Data

Data Source and Pre-Processing

- Source
  - NYC Dept. of Finance, 2017 properties rolling sales for all tax classes
  - Raw data set: five files (M, BX, B, Q and SI)
    [Click to link to Raw Data]
- Pre-Processing
  - All the usual data cleaning
  - For simplicity, all sales $< \$100,000$ filtered out
  - Grid granularity: Zipcode-wise ((*Lat.*, *Long.*) added to dataset)

Data Transformation

Boro-wise Normalization Normalize a quantity taken in zip code $\in$ *Boro* with respect to μ_{Boro} and σ_{Boro}. Essentially z-scores.

(Two examples next)

Transformation: Boro-wise Normalization (1/2)

Example 1 Normalized volume ZV in Brooklyn, January

Say Zipcode $= 11234 \in$ Brooklyn

$$ZV(Zip == 11234) = \frac{V(Zip == 11234) - \overline{V}_{\text{Bklyn., Jan.}}}{\sigma_{\text{Bklyn., Jan.}}}$$

Note

- $\overline{V}_{\text{Bklyn., Jan.}}, \sigma_{\text{Bklyn., Jan.}}$ from

$$\{V_{\text{Jan.}}(Zip_i) : Zip_i \in \text{Brooklyn}\}$$

- $ZV(Zip_i) == 0$: $V(Zip_i)$ at average level in Brooklyn, January
  - $\oplus$ $ZV(Zip_i)$: how $V(Zip_i)$ is above average
  - $\ominus$ $ZV(Zip_i)$: how $V(Zip_i)$ is below average

Transformation: Boro-wise Normalization (2/2)

Example 2 Normalized volume ZP in Manhattan, April

Say Zipcode $= 10023 \in$ Manhattan

$$ZP(Zip == 10023) = \frac{P(Zip == 10023) - \overline{P}_{\text{Manh., Apr.}}}{\sigma_{\text{Manh., Apr.}}}$$

Note

- $\overline{P}_{\text{Manh., Apr.}}, \sigma_{\text{Manh., Apr.}}$ from

$$\{P_{\text{Apr.}}(Zip_i) : Zip_i \in \text{Manhattan}\}$$

- $ZP(Zip_i) == 0$: $V(Zip_i)$ at average level in Manhattan, April
  - $\oplus$ $ZP(Zip_i)$: how $P(Zip_i)$ is above average
  - $\ominus$ $ZP(Zip_i)$: how $P(Zip_i)$ is below average

Boro-wise Normalization: Why? (1/2)

Philosophy: Everything is relative

1. Normalized volume *ZV*: sharp anomaly pattern *within* a borough

   Consider: $(\ldots, -1, 0, 1, \ldots)$ sharper constrast than $(\ldots, \mu - \sigma, \mu, \mu + \sigma, \ldots)$

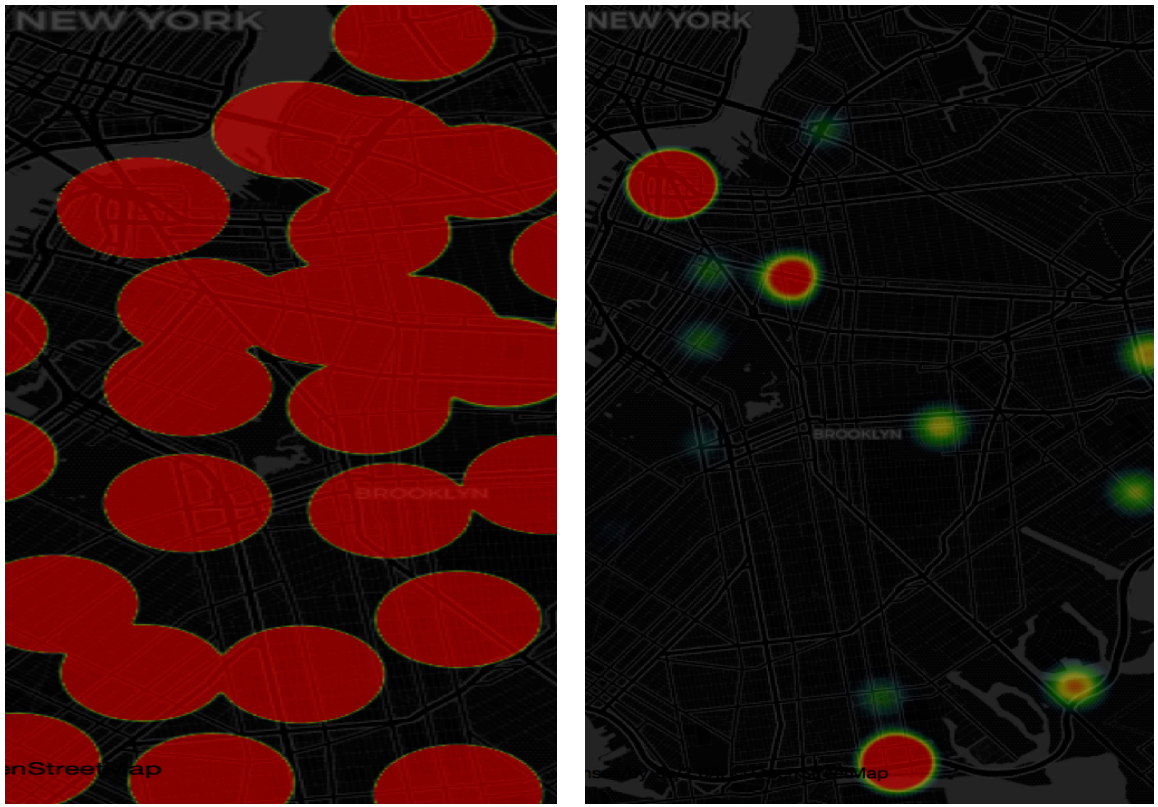
   (Example: next slide)

2. Normalized price *ZP*: fair price comparisons *across* boroughs
   - House price \$1M implies very differently in Manhattan than Staten Island
   - So does a \$100K fluctuation in Manhattan than Staten Island

Data Transformation: Why? (2/2)

Heat maps: House sales volumes, Brooklyn, May
(Heat map: intensity $\propto$ volume)

Left: Volume; Right: Normalized Volume



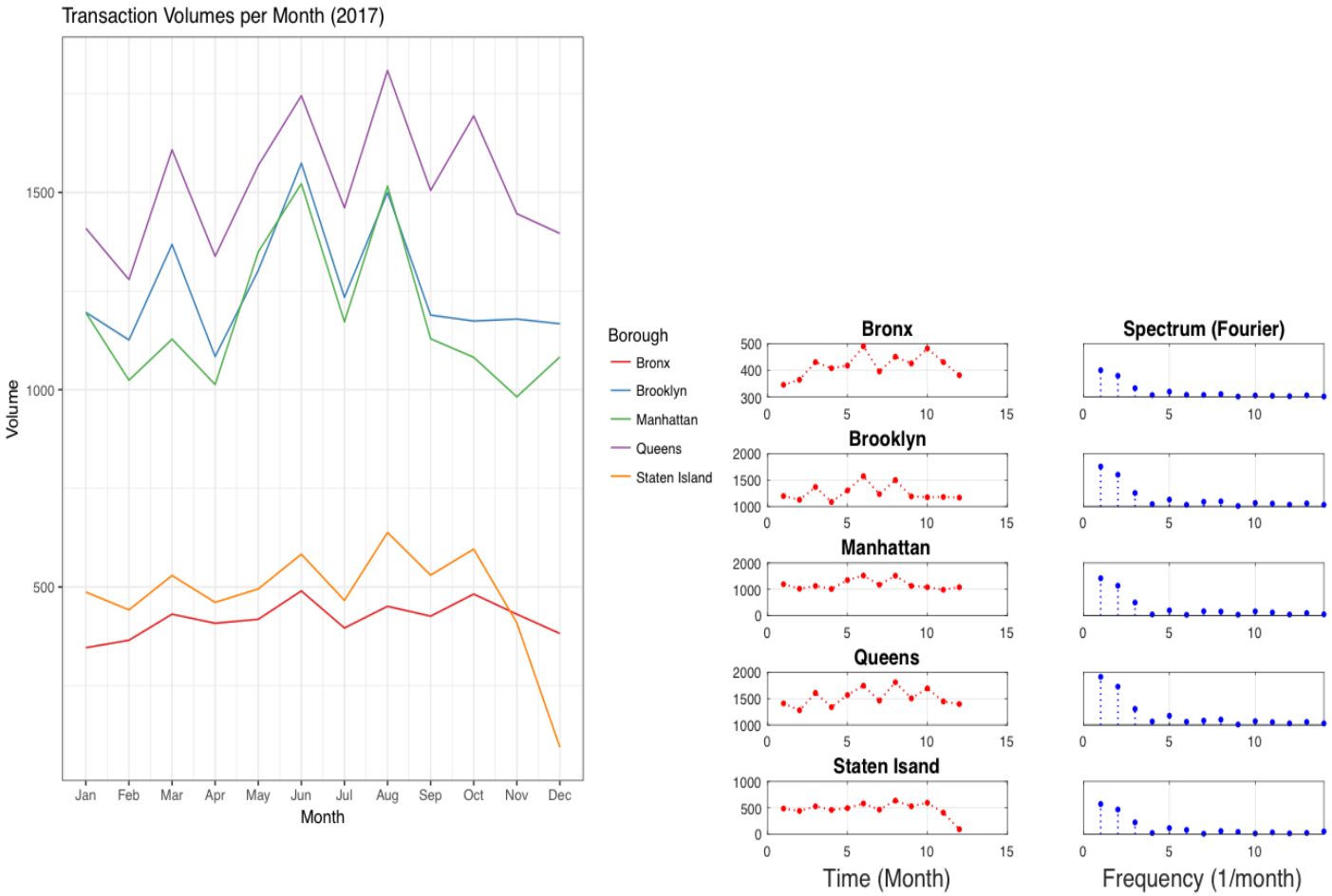
Part III. Shiny Demonstration

[Click to play on shinyapps.io]

Part IV. Additional Observations

Additional Observations (1/2)

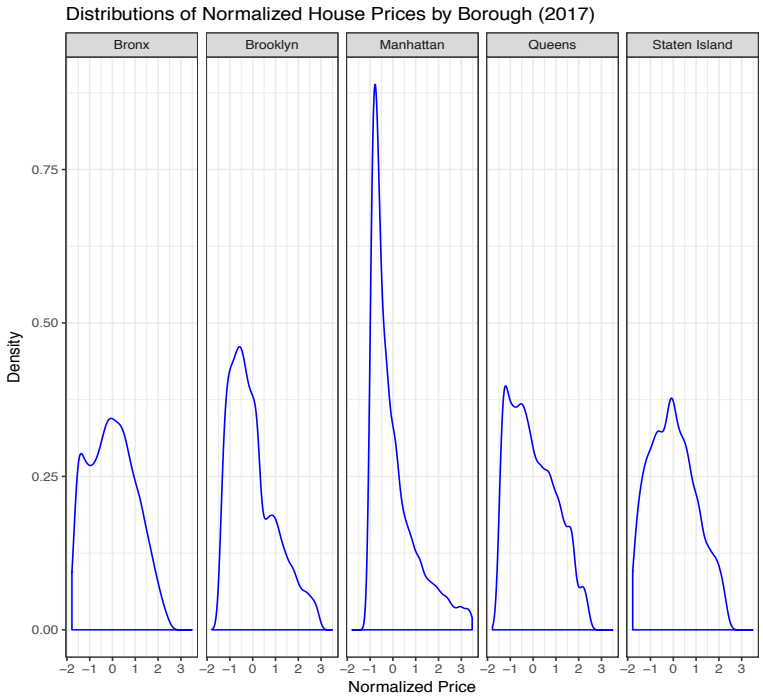
Time-behavior: Periodicity in transactions

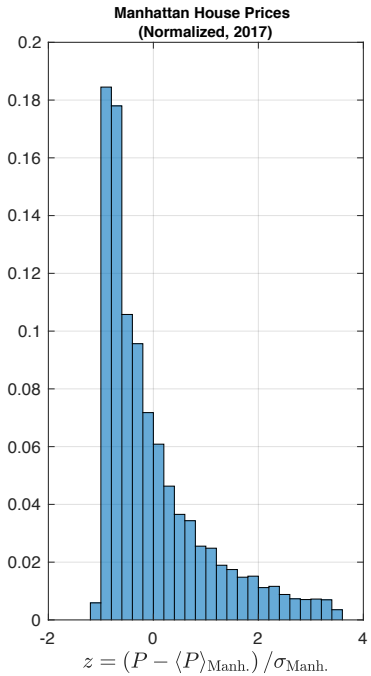


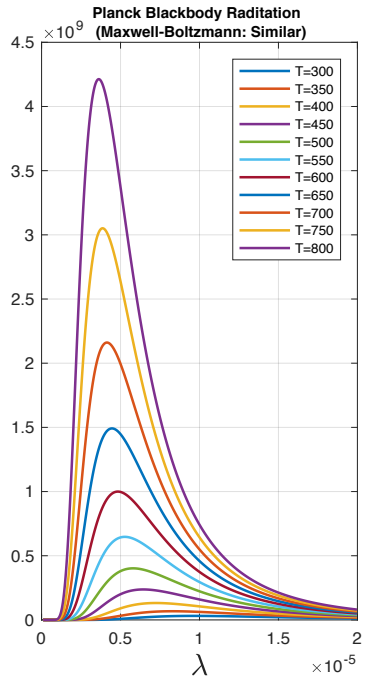
Additional Observations (2/2)

Postulate: In the thermodynamics of NYC house selling,

Temperature = Location







Thank You!

- Possible improvements:
  1. Refine plots and Shiny appearance
  2. Incorporate data from past years (2016 and prior)
  3. Refine study with different house categories
  4. . . .
- Questions?
- Thanks for the help with Shiny from AAron, Drace, Kathryn, and especially, Zeyu
- Thank you!