Notes On Shiny Project: NYC House Sales 2017

Zhengqing (James) Chen

January 31, 2018

1 Data Source

NYC Department of Finance, Rolling Sales (last 12 months, January 2017 to December 2017) for all tax classes, sorted by borough and neighborhood. The data set consists of five files corresponding to the five boroughs of NYC: Manhattan, Bronx, Brooklyn, Queens and Staten Island. For raw data, see

`http://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page`

2 Data Pre-Proccessing

Beside the usual data cleaning, for simplicity of discussion, all sale records with dollar amount less than $\$100,000$ were filtered out.

3 Data Transformation

In this mini project, the focus is on visualizing the transaction volumes, henceforth shortened as "volume". We also did an observation of price.

Volume Simply the number of houses sold, not the sales dollar amount. It is always discussed within specific context. For example, we analyzed monthly volumes per zip code. That is, the number of houses sold within each zip code in the twelve months of 2017.

Perhaps the most important data transformation in this project is borough-wide normalizations, essentially $z-$scores.

Borough-wide Normalization Normalize an interested quantity with respect to μ and σ of that quantity over a borough where that quantity was taken.

See the following two examples:

a. Normalized volume in Brooklyn in January, ZV. Say the zip code is $11234 \in$ Brooklyn, we have

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

where $\overline{V}_{\text{Bklyn., Jan.}}$ and $\sigma_{\text{Bklyn., Jan.}}$ are calculated from all

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

Note that a zero ZV value at a particular zip code $\text{Zip}_i \in \text{Boro}_j$ does *not* mean there was no house sold at Zip_i, but that it is at the average level of Boro_j. Likewise, a positive/negative ZV at Zip_i indicates how it is above/below average in Boro_j.

b. Normalized house price in Manhattan in April, ZP. Say the zip code is $10023 \in$ Manhattan, we have

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

where $\overline{P}_{\text{Manh., Apr.}}$ and $\sigma_{\text{Manh., Apr.}}$ are calculated from all

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

Note that a zero ZP at zip code $\text{Zip}_i \in \text{Boro}_j$ means it is at the average level of Boro_j, and a positive/negative ZP indicates how it is above/below average in Boro_j.

Why we do such transforms?

a. The major benefit of introducing normalized volume is to make the anomaly pattern of transactions within a borough sharper. For example, on a heat map where the heat intensity is set to be the volume, the contrast of values of say $(\ldots, \mu - \sigma, \mu, \mu + \sigma, \ldots)$ is less sharp than their normalized values $(\ldots, -1, 0, 1, \ldots)$.

b. The major benefit of introducing normalized price to make the price comparisons across boroughs fair. For example, a house price of \$1 million implies very differently in Manhattan than in Staten Island, so does a fluctuation of $\$100,000$ in Manhattan than in Staten Island.

See the Shiny App demonstrations for the above two points. Behind all these is the philosophy that absolute magnitudes themselves reveals little; everything is relative and needs to be considered in perspective.

4 Other Technical Notes

Periodic fluctuation of volumes is examined via its Fourier analysis, FFT (fast Fourier transform). Distributions of normalized prices over five boroughs are drawn analogy to be Planck or Maxwell-Boltzmann like, where the former is for "gases" of photons and the latter ideal gas in thermal equilibrium when states depend on temperature only. See the Shiny App demonstration of the above two points.