DGMD S-76 Individual Report
Anna Romanova

Using of unsupervised learning approaches to clarify missing price data in The one-year (2016-2017) collection of property sales in NYC from the City of New York.

*image [5]

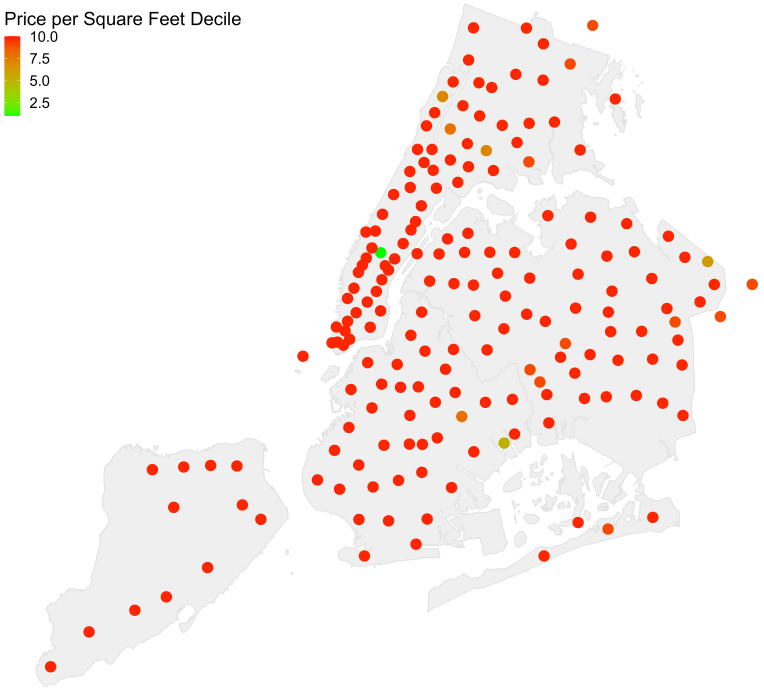
Preface

In this report some of the unsupervised learning approaches are used to clarify missing price data in The one-year (2016-2017) collection of property sales in NYC from the City of New York [1]. Preliminary cleaning and reformatting of the data includes: conversion to numeric format and replacement of undefined values with corresponding averages of those attributes or zero, adding latitude and longitude by zip code from [2], conversion[3] to NYC map[4].

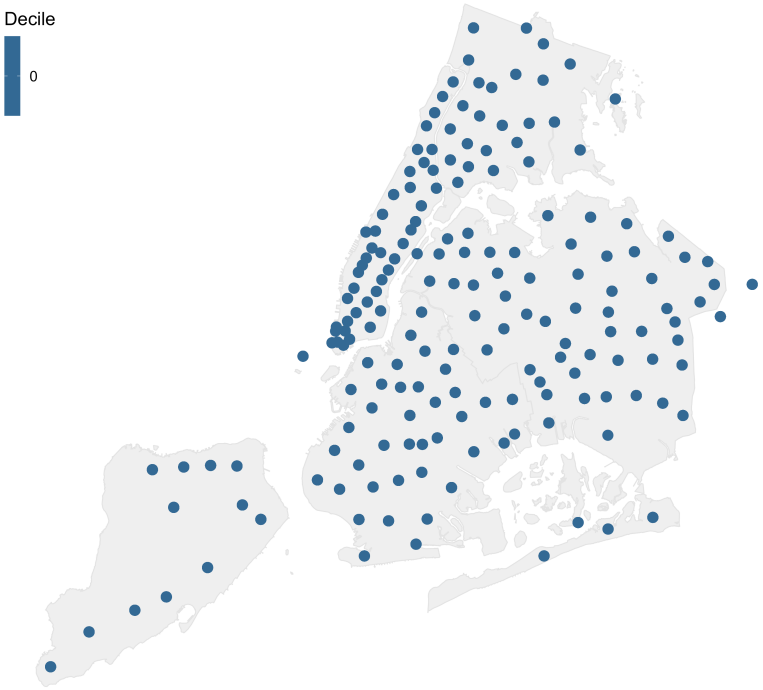
Sales Price Data

The data have 84548 observations from which 82410 can be located on NYC map. 49710 have price data and the other 32700 do not (even if in the same building). Priced and not priced objects are shown on NYC counties map giving the general understanding in what Decile of Price per Gross Square Feet the unknown object could be (based on the most common price level in the district).

Price per Square Feet Decile



Price per Square Feet Unknown

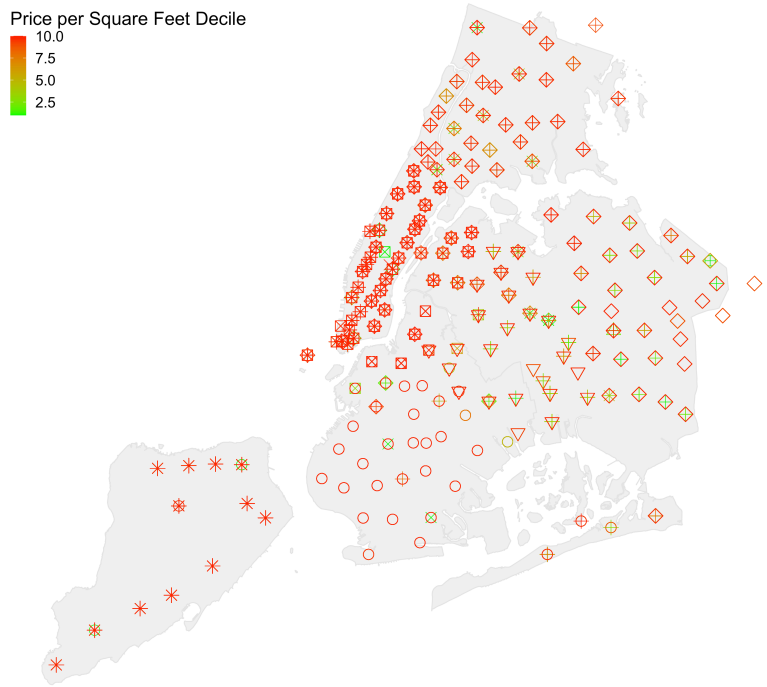


While some objects are in different Deciles of Price per Square Feet, most of them are on the top Deciles, so we add clustering principal component analysis.

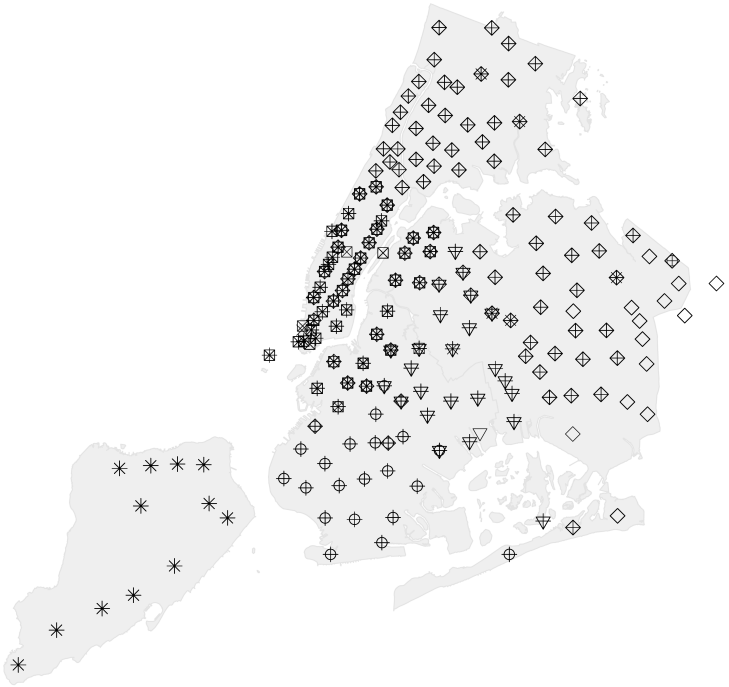
K-means clusters and Principal Component Analysis

Function `kmeans` performs K-means clustering on *explicitly scaled* (e.g. `kmeans(scale(x),2)`) property sales in NYC data for 10 clusters: `cluster` attribute in the output of `kmeans` indicates cluster membership (shown by shape in the plot). Variables for clustering are: Residential Units, Commercial Units, Total Units, Land Square Feet, Gross Square Feet, Year Built, Longitude, Latitude. The same variables are used to predict cluster for objects with unknown price.

Price per Square Feet Decile and Cluster



Predicted Cluster for Objects with Unknown Price



There are several visible clusters of objects by their attributes (in Manhattan, in the Staten Island, in the south of Brooklyn and the south of Queens, in the north of Bronx, etc.). Using the predicted cluster for the objects without price it is possible to look to nearby objects from the same cluster and get the understanding of the possible price (for example, around JFK we may see the objects from the same clusters with price and without price).

The most of differences between the objects (53.41%) are explained by two main components (principal components):

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7      PC8
Standard deviation     1.6750 1.2114 1.0906 1.0045 0.84075 0.72508 0.54399 0.004254
Proportion of Variance 0.3507 0.1834 0.1487 0.1261 0.08836 0.06572 0.03699 0.000000
Cumulative Proportion  0.3507 0.5341 0.6828 0.8089 0.89729 0.96301 1.00000 1.000000
```

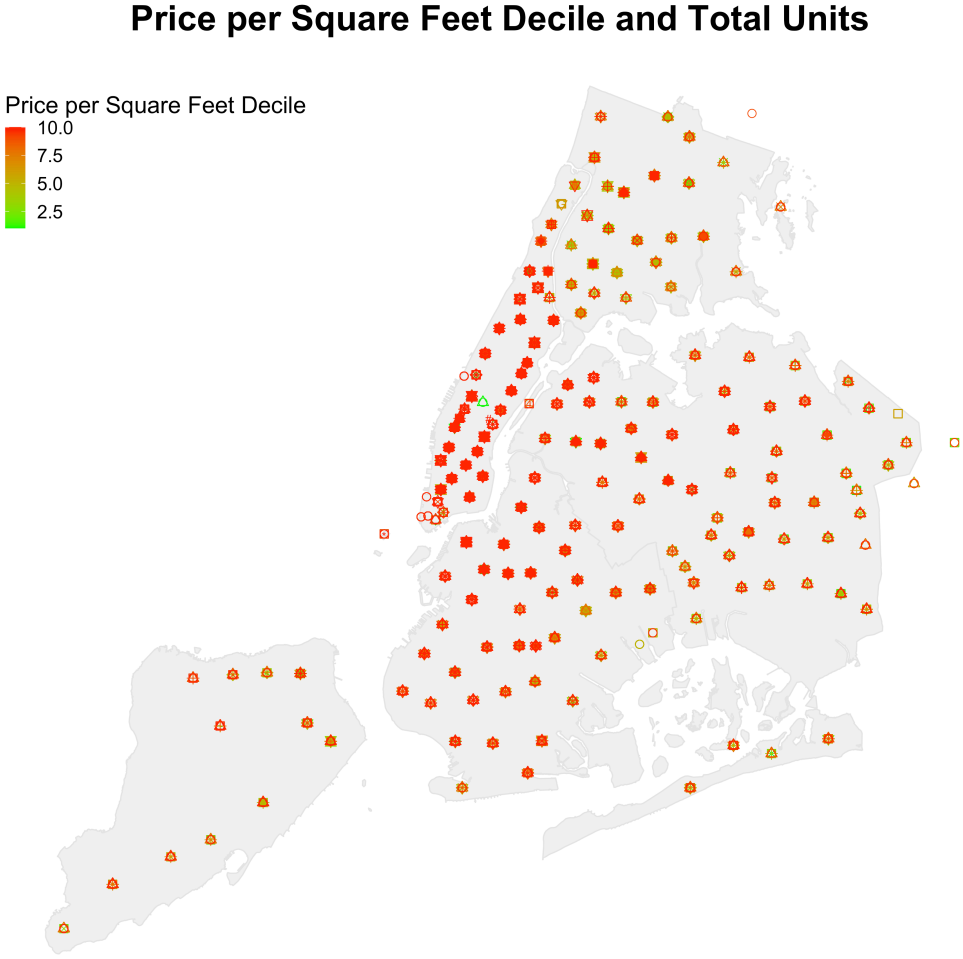
The variable that contributes the most to the Dimensions 1, 2 of PCA is Total Units, and the least contributing variable to Dimensions 1, 2 of PCA is Year Built:

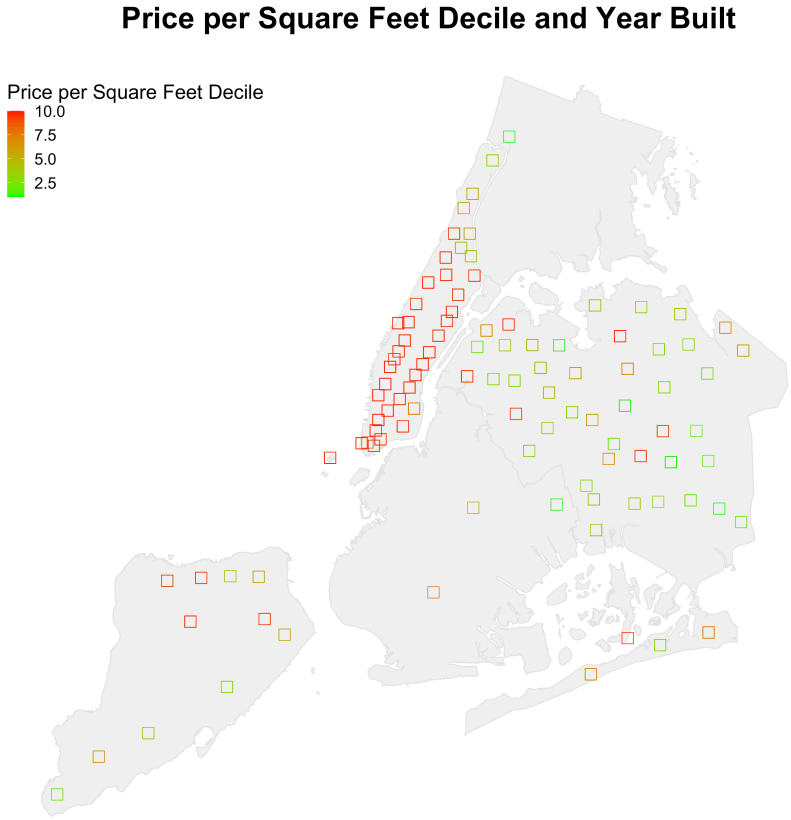
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
RESIDENTIAL.UNITS	27.299775843	0.000003965012	2.9701112	0.223516120	21.5217777668	0.513135034
COMMERCIAL.UNITS	4.220309969	0.002590024497	67.6002673	0.627565641	9.1452334228	0.147645052
TOTAL.UNITS	29.452341713	0.000853173077	10.7494246	0.005046798	4.2325207823	0.131604863
LAND.SQUARE.FEET	13.446077148	0.229876178093	11.1488585	0.190571197	63.9915438370	0.003236991
GROSS.SQUARE.FEET	25.512758533	0.005034294732	7.0338227	0.017993842	0.0008408136	0.338726824
YEAR.BUILT	0.002961718	1.431821068053	0.4046416	94.612004524	0.4058659939	3.128132740

	Dim.7	Dim.8
RESIDENTIAL.UNITS	14.08958409	33.3820959485426
COMMERCIAL.UNITS	1.93756268	16.3188259596738
TOTAL.UNITS	5.12914630	50.2990617383349
LAND.SQUARE.FEET	10.98983614	0.0000000010558
GROSS.SQUARE.FEET	67.09082281	0.0000001760132
YEAR.BUILT	0.01456088	0.0000114733690

No significant contribution is shown from Latitude and Longitude, suggesting that ranges are too small to produce noticeable impact to PCA.

Total Units is the most contributing attribute to the variability of the objects sold in NYC in 2016 - 2017: clear patterns between Total Units (shown by shape) and Price per Square Feet (shown by colour) are visible on Staten Island, in the south of Brooklyn, south of Bronx. Different combinations of shapes (Total Units) and colours (Price per Square Feet) suggest that in different districts Total Units priced differently (i.e. in Manhattan any types of Total Units are top priced).





As it is shown in the principal component analysis Year Built (shown by shape) does not contribute significantly to variability of the objects.

While general analysis shows main components, influencing real estate prices in NYC in 2016 - 2017, further analysis may be advised in dimensions of specific interest: specific Decile of price, specific district, specific residential or commercial units, etc.

References:

[1] The one-year (2016-2017) collection of property sales in NYC from the City of New York. https://www.kaggle.com/new-york-city/nyc-property-sales.

[2] US Zip Code Latitude and Longitude. https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/.

[3] Mapping in R using the ggplot2 package. 2014. http://zevross.com/blog/2014/07/16/mapping-in-r-using-the-ggplot2-package/.

[4] NYC Department of City Planning. https://www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page.

[5] https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City