University of Illinois System

Junseok Yang
Himanshu Kumar
Matthew Jewell

Beijing Housing Prices

Predicting Total Price and Identifying Over- and Underpriced Listings

Data Overview

- 318,851 overservations
- 26 variables
- Challenges: Missingness and translation
  - “DOM” had nearly 50% of observations missing
  - Translation and processing issues with “Floor” variable
- Limitations: Mostly computational

UNIVERSITY OF ILLINOIS SYSTEM

Project Objectives

1. Regression: Create a regression model which can accurately predict a real estate listing's total price, given other information about the listing.

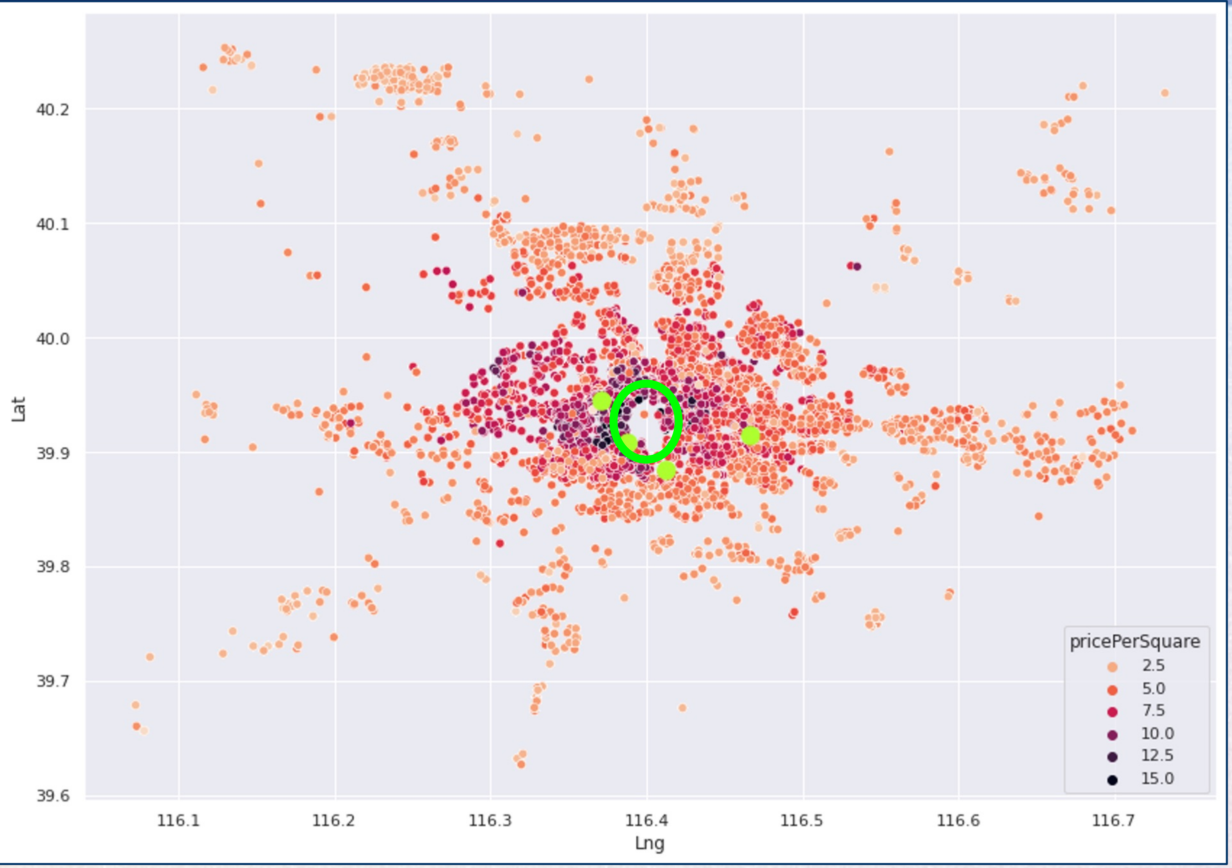
1. Clustering: Create clusters which effectively group listings with similar attributes to identify over- and underpriced listings.

UNIVERSITY OF ILLINOIS SYSTEM

Setting a Benchmark: Other Groups' RMSEs

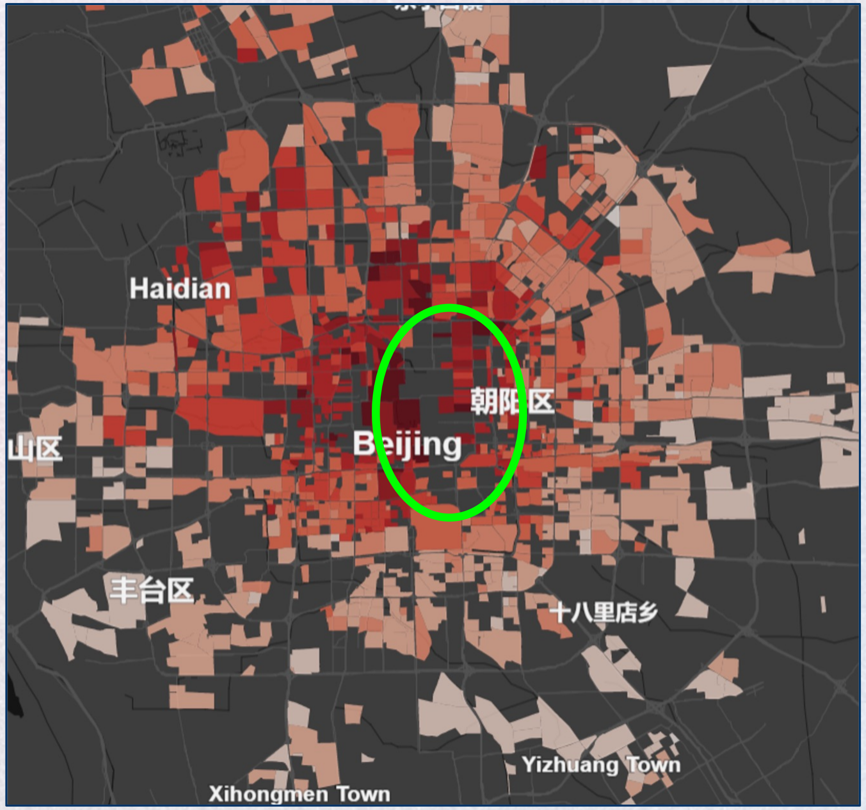
RMSE by Group Predicting Total Price				
A	B	C	D	E
124.3104	144.755	136.775	89.822	126.304

- Many groups used metrics other than RMSE to measure model performance, including RMSLE and score functions from packages in Python.
- Our group opted for RMSE because it is easily recognized and understood.



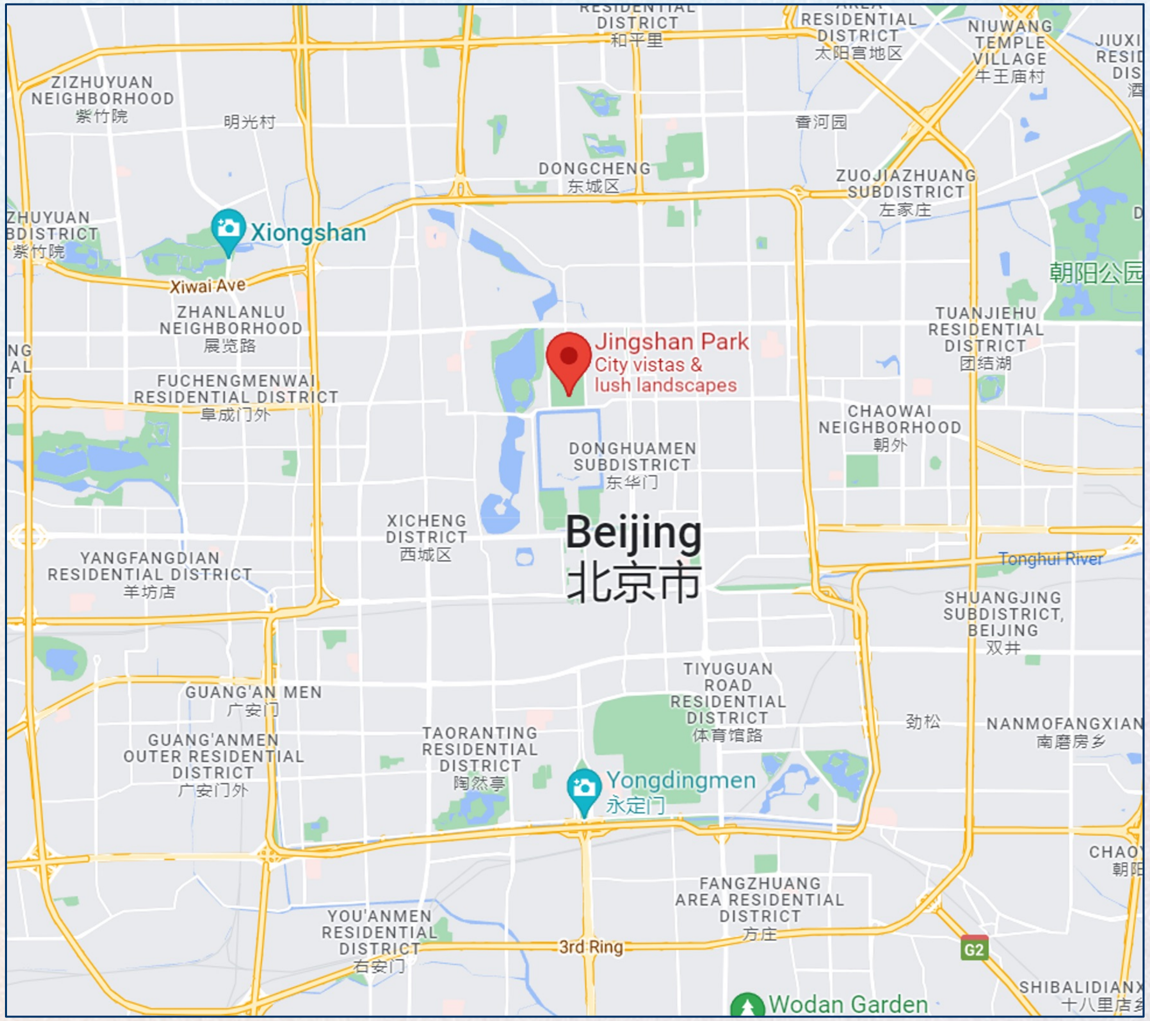
Literature Review Findings

Latitude and longitude data fit into Beijing city-blocks using spatial joining. The colors indicate average property price per square meter.



Map of Beijing with Tiananmen Square, Temple of Heaven, Xin Jiekou, and Beijing International Trade Center (green dots).

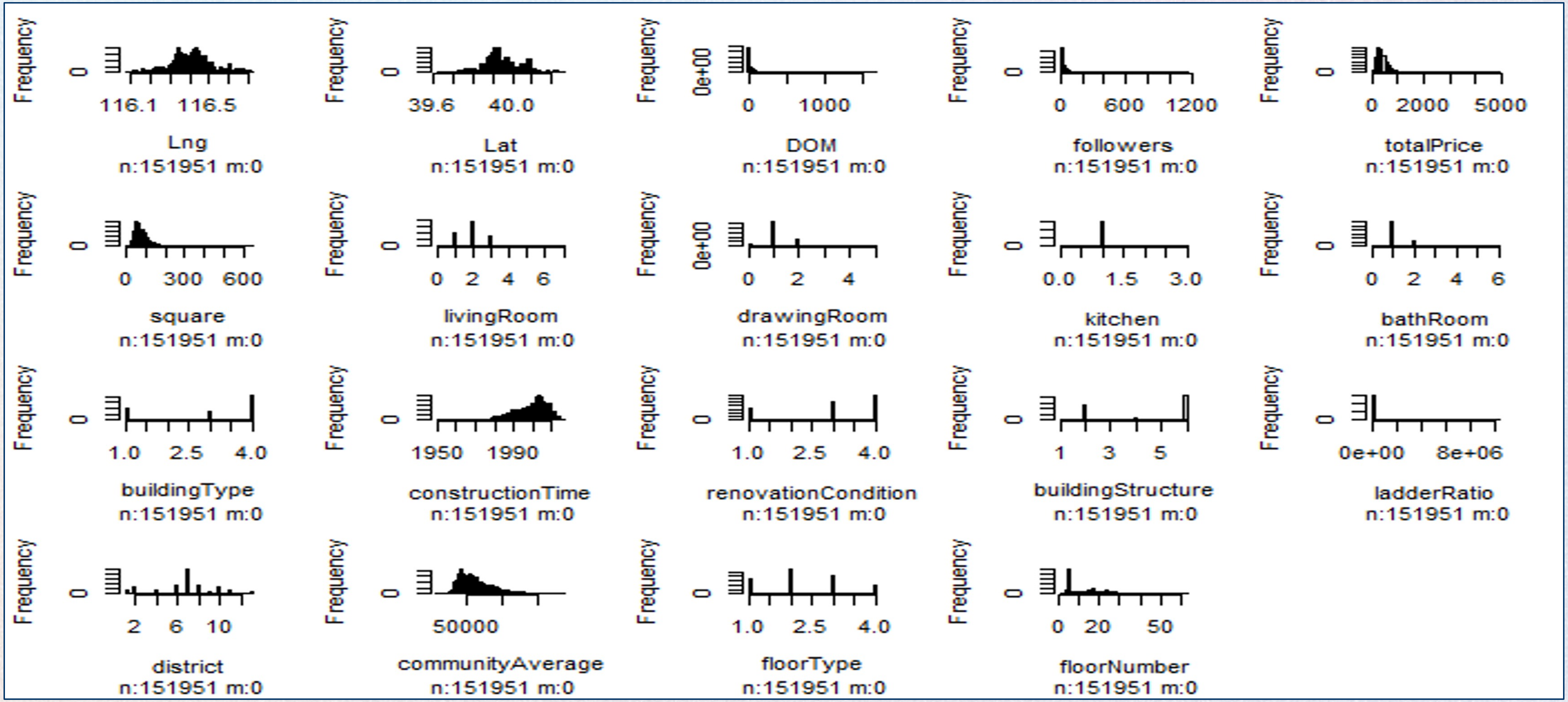
Identifying the Epicenter



- Examining the images on the previous slide, we noted what appeared to be an epicenter from which the high listing prices radiated.
- Using Google Maps, we were able to identify a landmark in that region: Jingshan Park.
- We found the latitude and longitude for Jingshan Park and used that information to create a "Distance" variable which replaced our longitude and latitude variables with one measurement of how far a listing was from Jingshan Park while accounting for the Earth's curvature.[1]

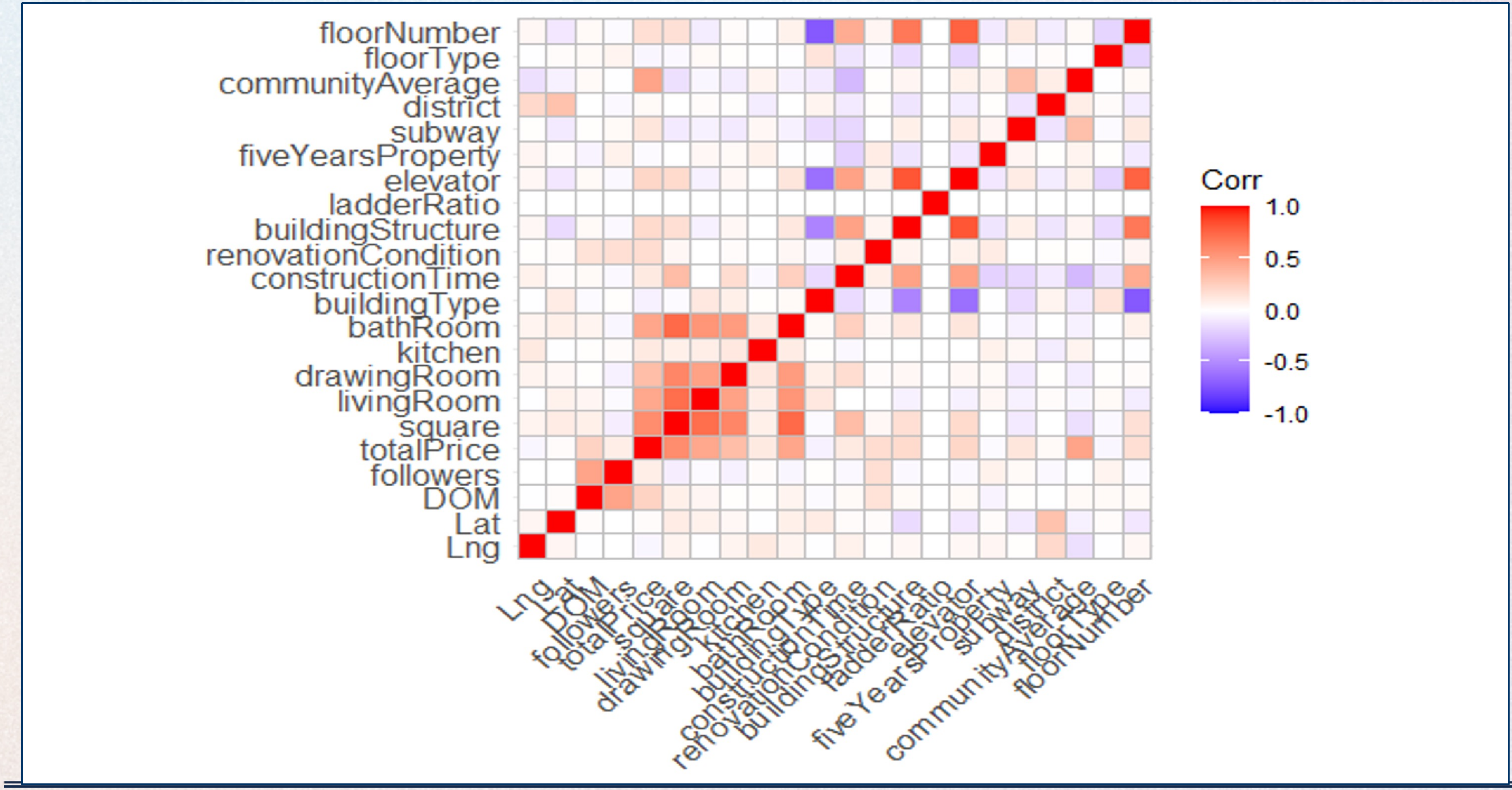
[1] This was accomplished with a function and the "geosphere" package.

HISTOGRAM OF VARIABLES



UNIVERSITY OF ILLINOIS SYSTEM

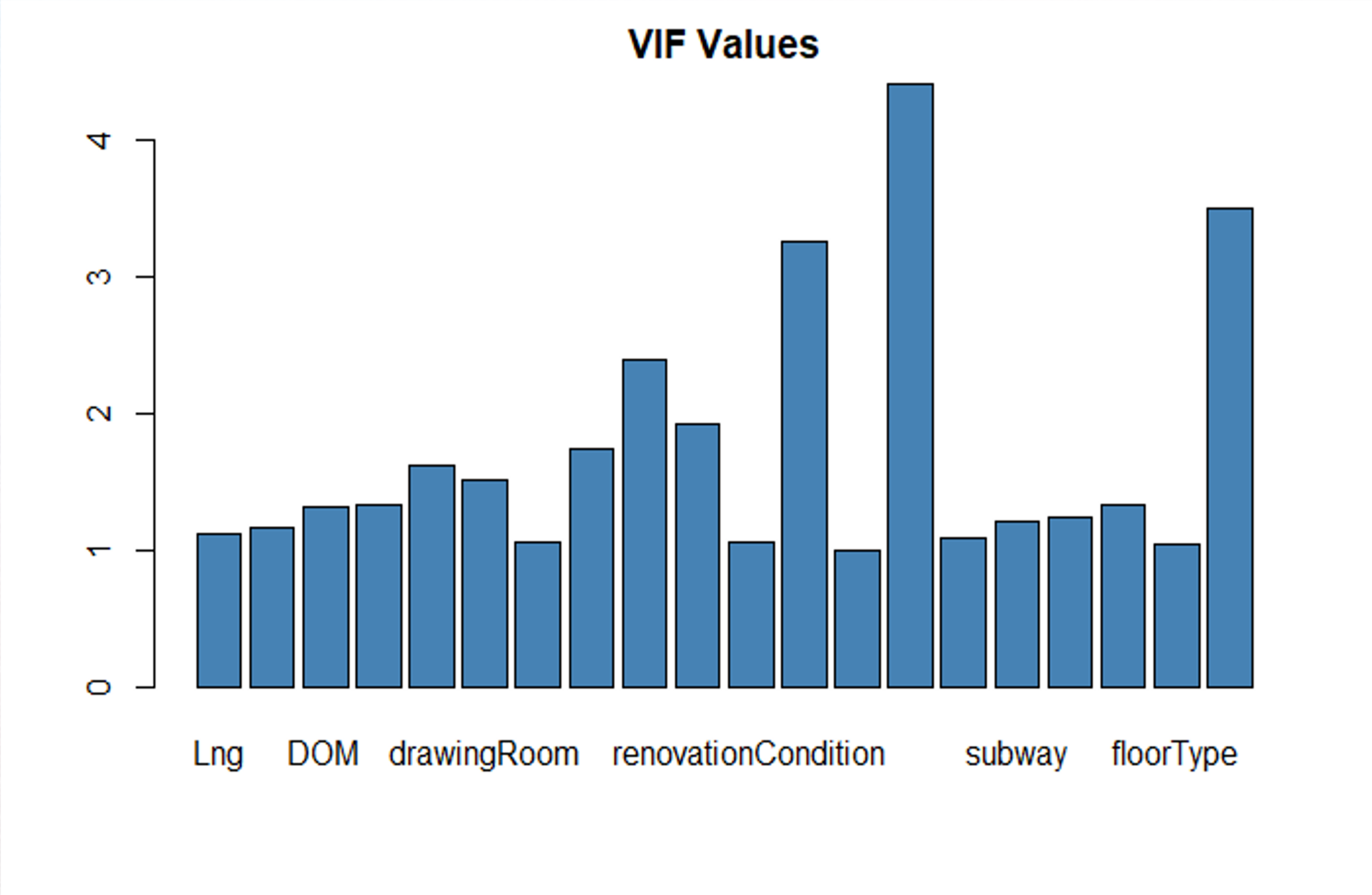
VARIABLES CORRELATION HEATMAP



UNIVERSITY OF ILLINOIS SYSTEM

VIF VALUES

VIF value of floor type = 1.042120

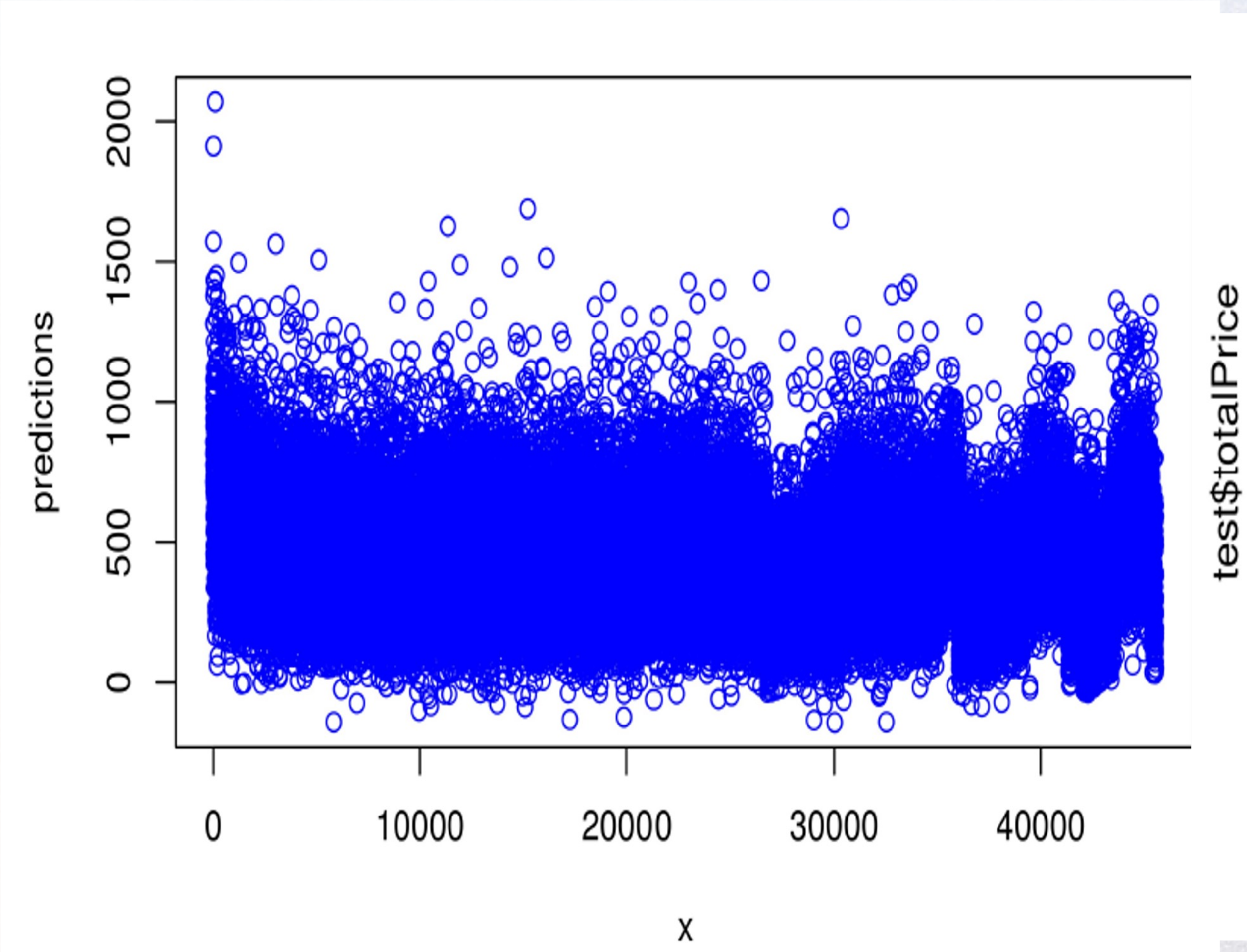


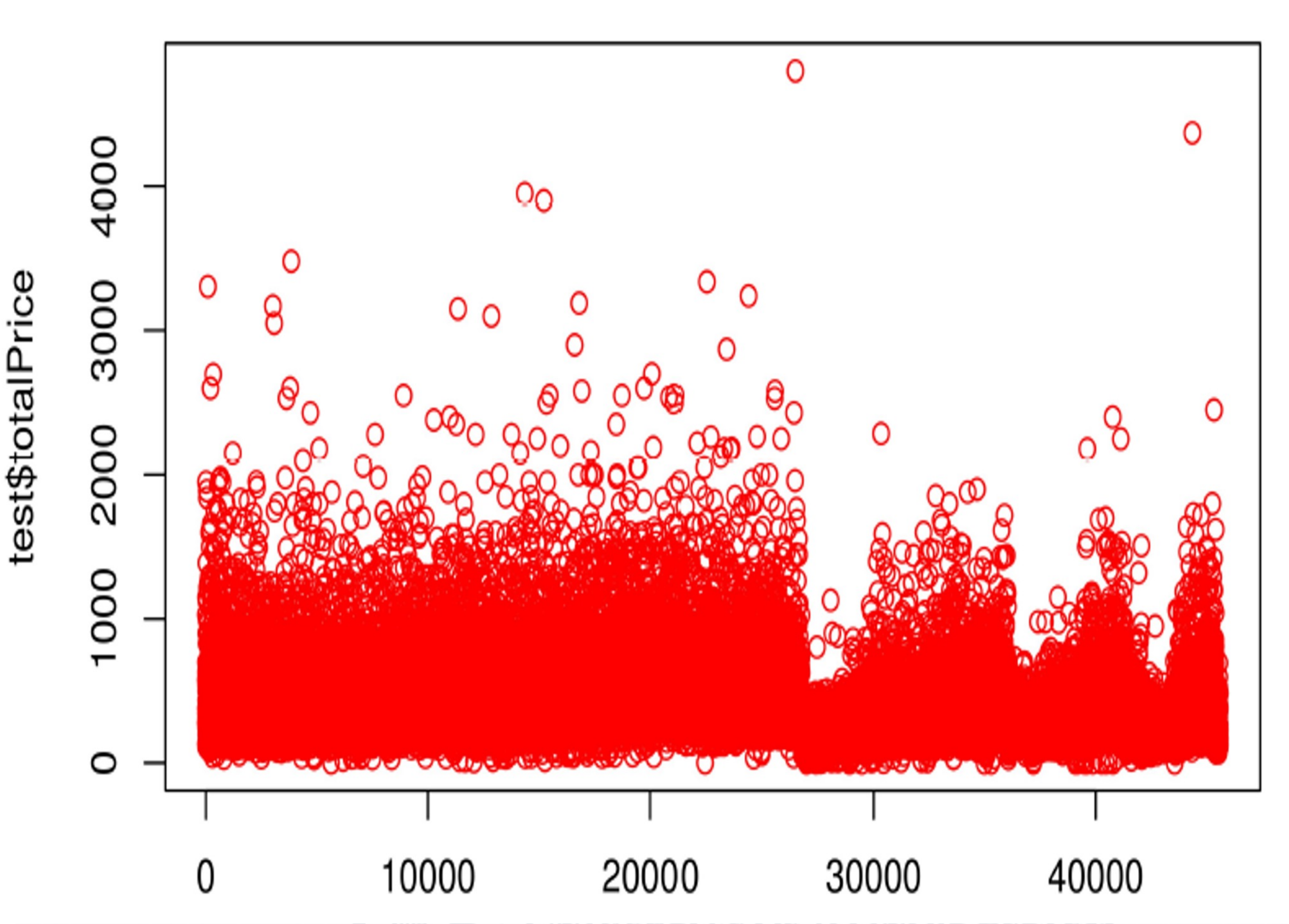
UNIVERSITY OF ILLINOIS SYSTEM

LINEAR REGRESSION

- A model was selected using the step function and based on the least AIC value (1074655). According to this, Ladder ratio and Longitude were not statistically significant.
- Based on lowest Mallows' Cp using the regsubsets function, the best model was the one that excluded Longitude and Ladder Ratio. (Mallows' Cp: 20.07409)
- Ridge and Lasso regression models were selected using the cv.glmnet function with nfolds = 10.
- The minimum lambda values for ridge and lasso are 0.0083 and 0.0085 respectively.

PREDICTED VS. ACTUAL VALUES (LINEAR REGRESSION)

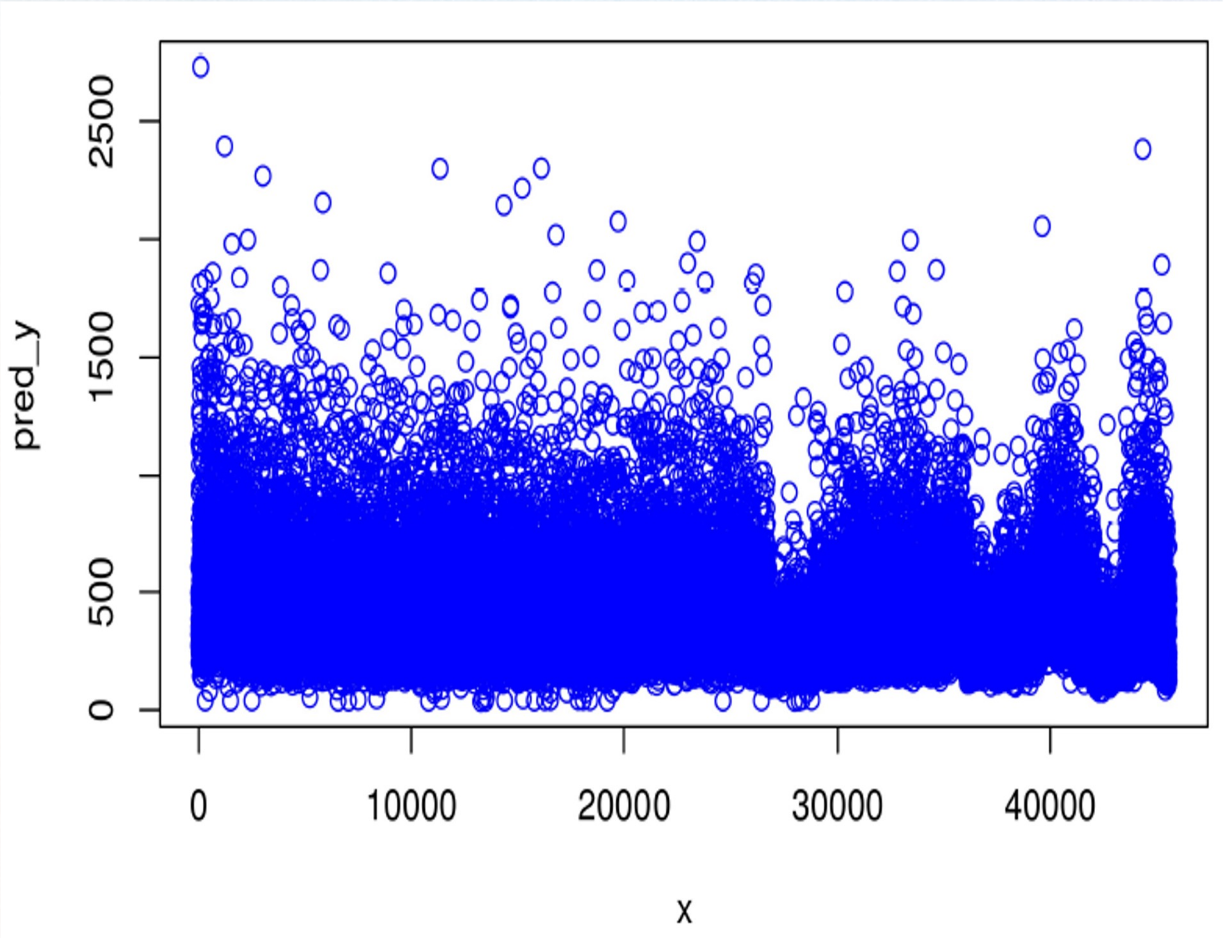


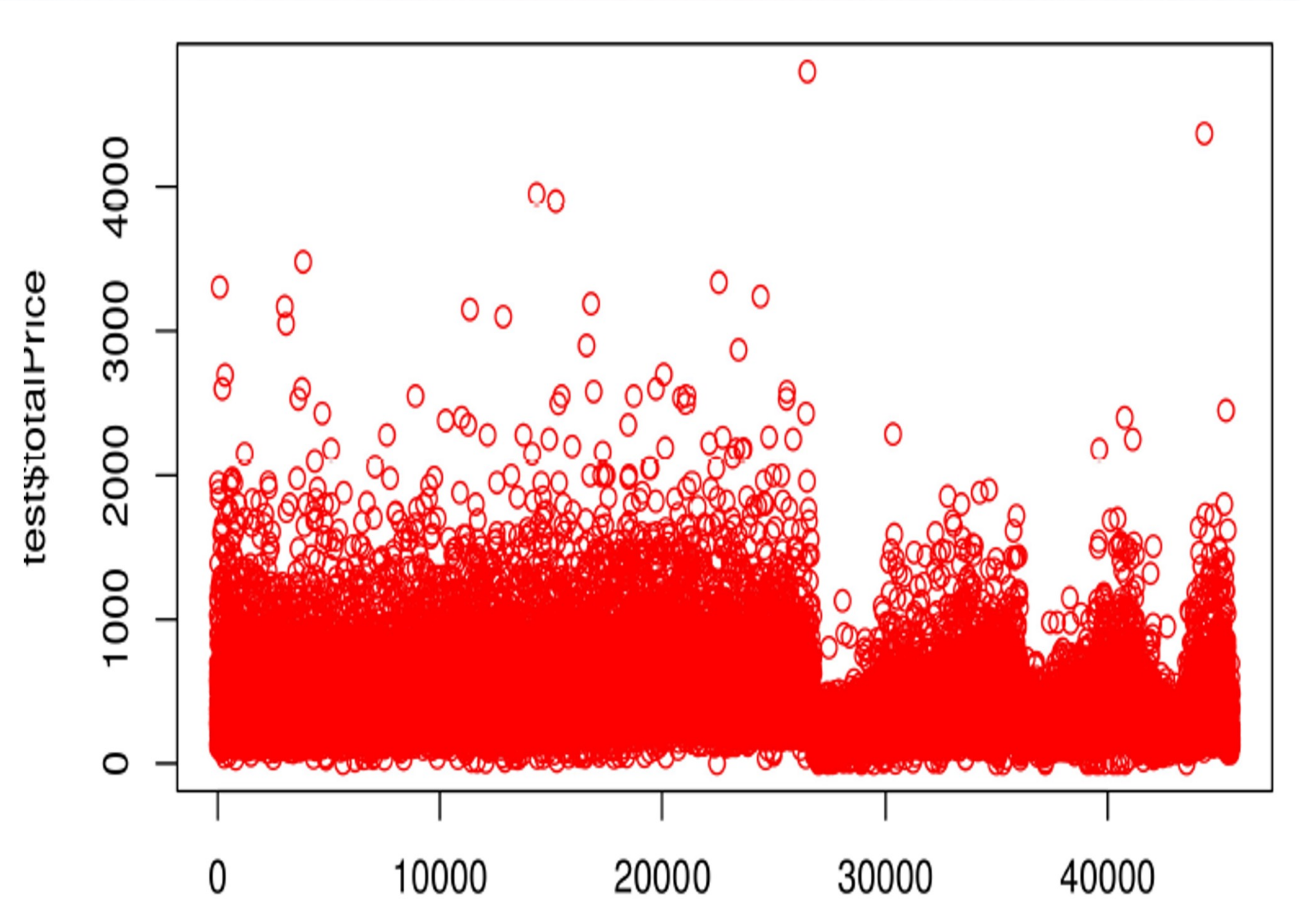


K-Nearest Neighbors

- KNN regression was done using knnreg function.
- The tuning parameter was the k value.
- The ideal k value, based on the least RMSE value, was found to be k = 8.
- The RMSE for KNN regression is less than linear regression and the R-squared value is also greater for the KNN model.

PREDICTED VS. ACTUAL VALUES (KNN)



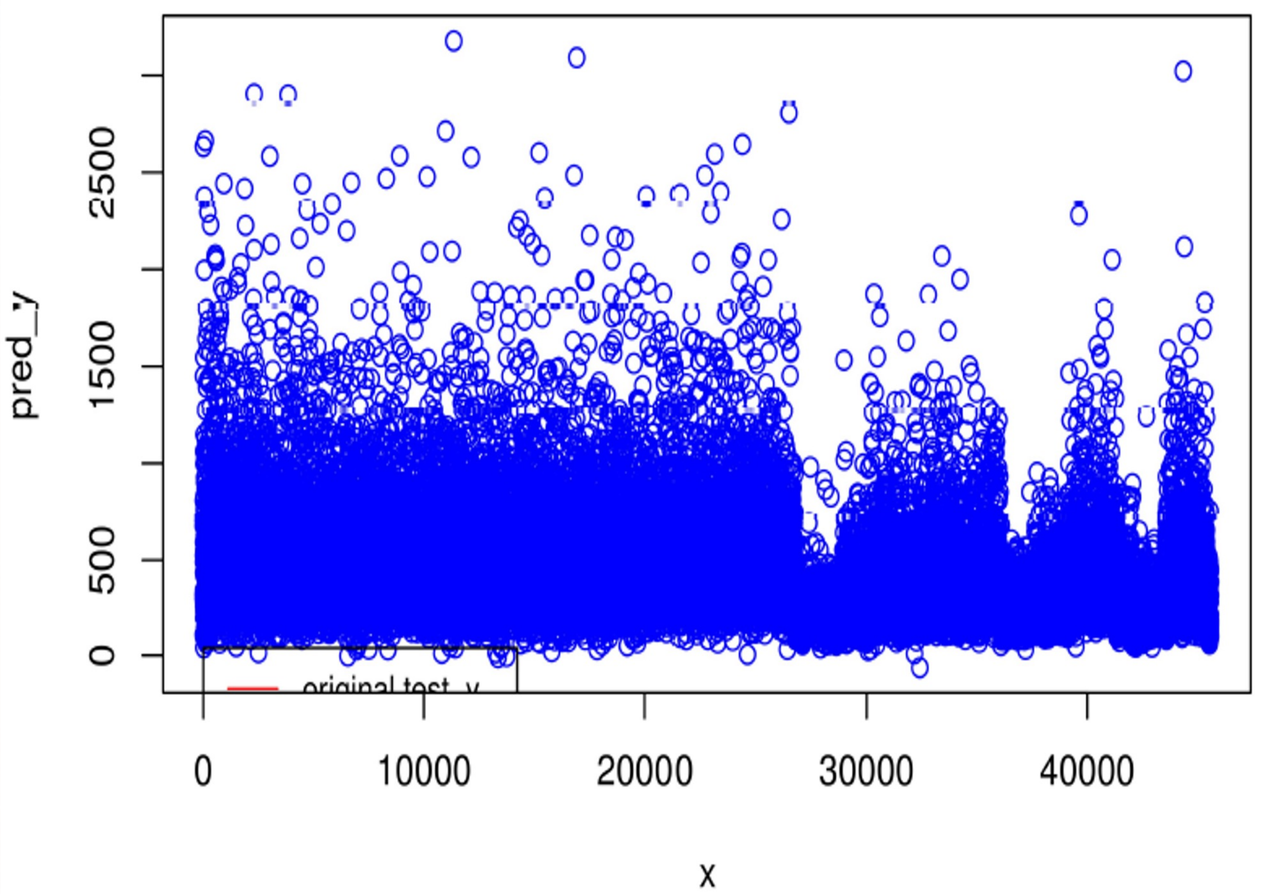


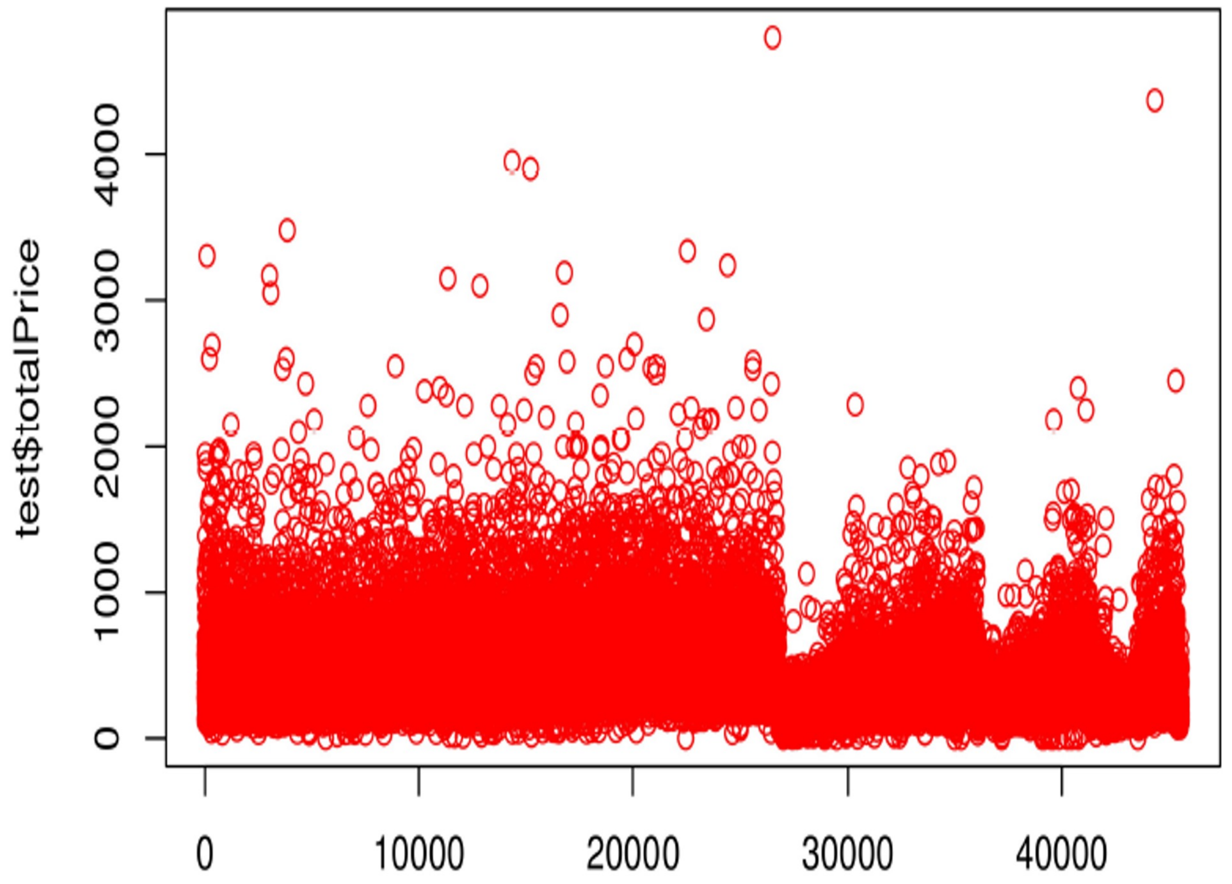
Extreme Gradient Boosting

- Used Caret package to train the model using xgb.train.
- The tuning parameters include “max.depth”, “nrounds”, and “lambda”.
- The ideal values for these parameters were found to be 7; 400 and 3.
- The least values for test-RMSE as well as the train-RMSE were found to be at the 257th round.

UNIVERSITY OF ILLINOIS SYSTEM

PREDICTED VS. ACTUAL VALUES (XGBOOST)





UNIVERSITY OF ILLINOIS SYSTEM

COMPARING THE DIFFERENT REGRESSION MODELS

Regression Model	Training RMSE	Testing RMSE	R^2 Value	Tuning Parameters
Linear Regression	158.076026	157.3966963	61.34%	N/A
Ridge	158.076026	157.396601	61.34%	lambda= 0.0083
Lasso	158.076026	157.396601	61.34%	lambda= 0.0085
k-nn	121.033	137.8368601	70.67%	k = 8
XgBoost	56.81966	91.91300234	86.95%	max.depth = 7 nrounds = 400 lambda = 3

UNIVERSITY OF ILLINOIS SYSTEM

Unsupervised Learning

Cluster Analysis

Purpose / Research Goal

- Group / Identify the data
  - Similar / dissimilar
  - Notable features that share within or distinguish between the groups

- Meaningful trends behind the real estate market
  - Some common beliefs…
  - Association between ’price’ and top 10 predictors

Method / Data Pre-processing / Limitation

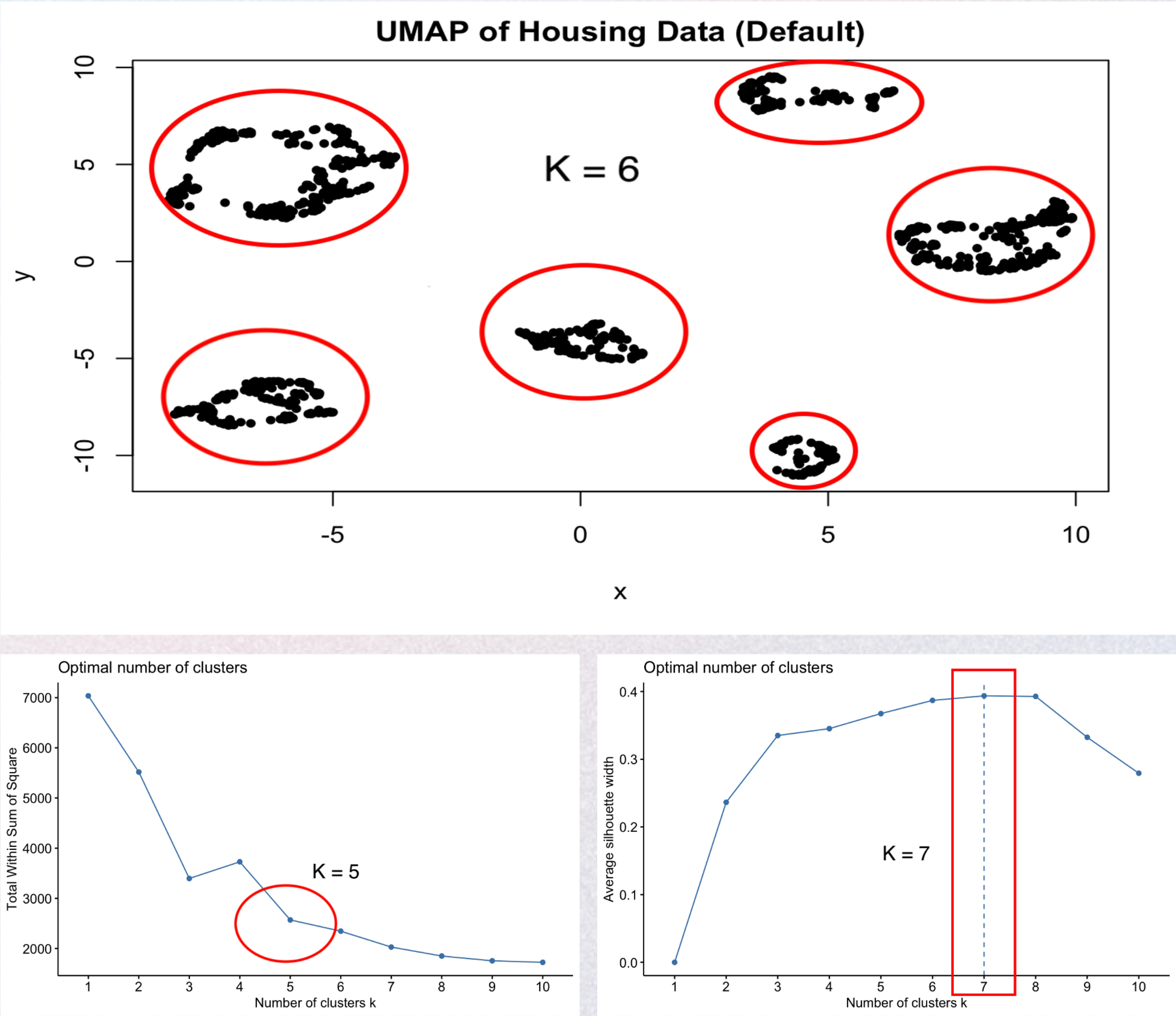
- Method
  - K-Means Clustering
- Data Pre-processing
  - Random sample size of 1000
  - Scaling & Conversion
    - Mean 0 & Standard Deviation 1
    - Gower's distance
      - Convert mixed data (Numeric + Non-numeric) to numeric
- Limitation
  - Representativeness
  - Lack diversity

⋮

Optimal Number of Clusters

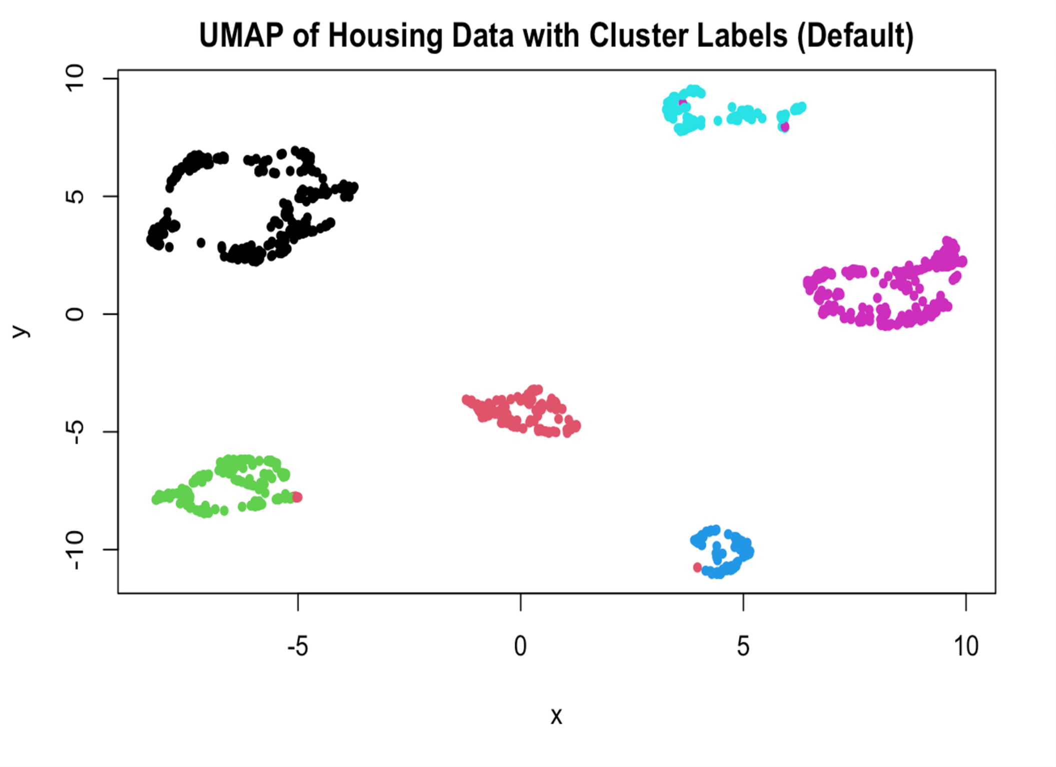
- UMAP (Uniform Manifold Approximation and Projection)
  - Dimensionality Reduction
  - Data formation
- Elbow method
  - Minimize within cluster variation
- Silhouette score
  - Cohesiveness within a cluster & Separation between clusters

➢ K = 6

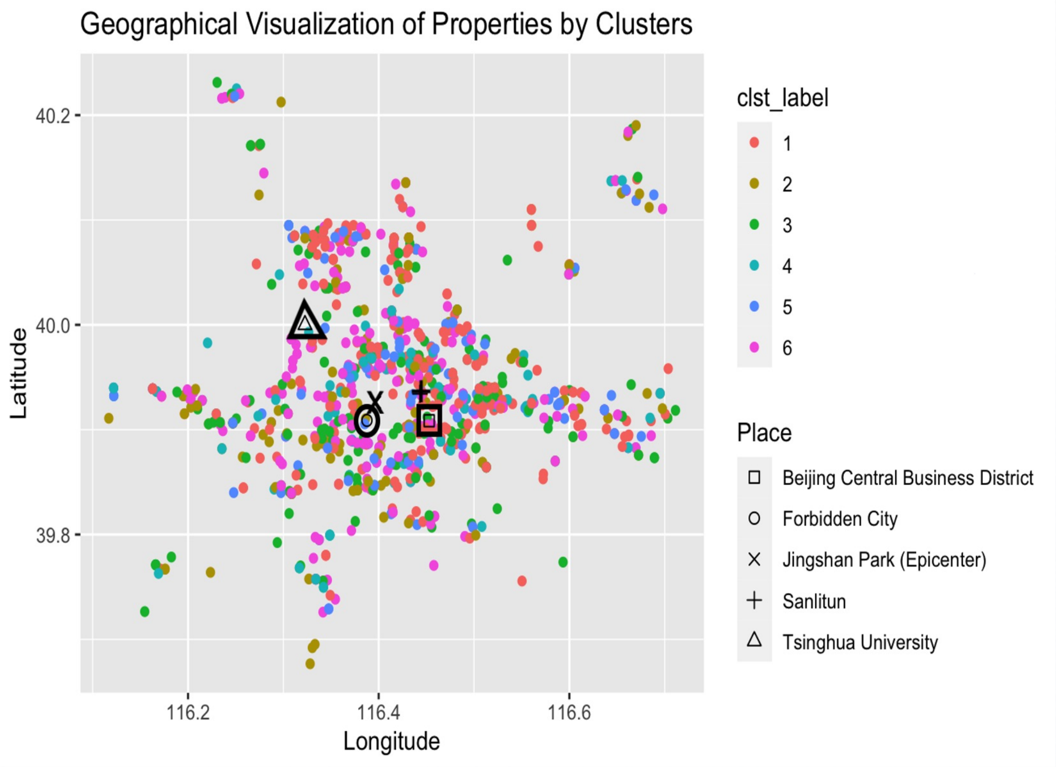


Cluster Analysis

- **Match with the groups**



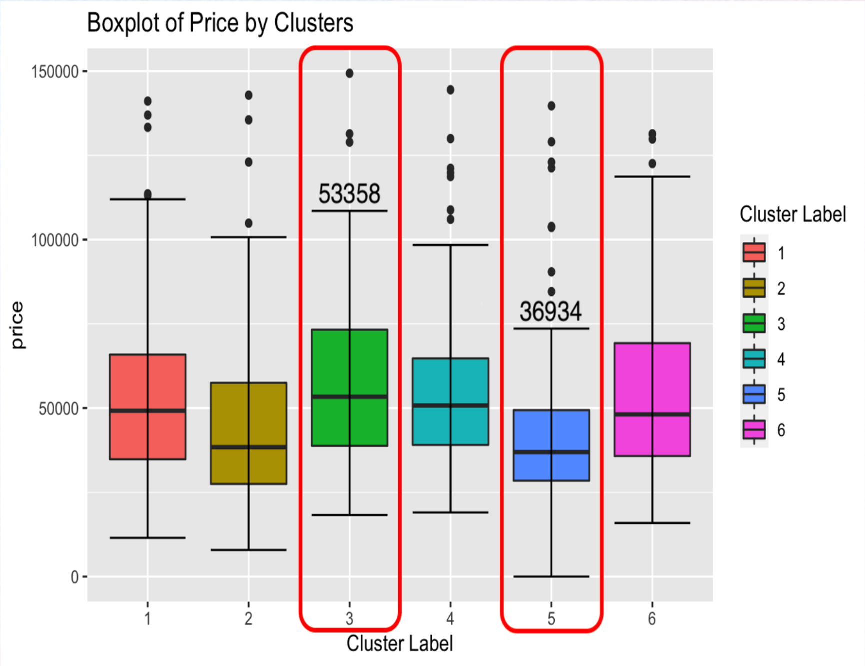
- **Not meaningful trend based on geographics**

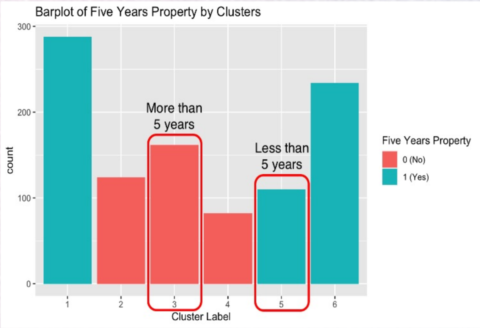


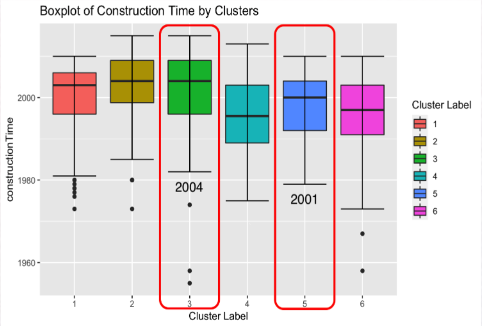
Cluster Analysis

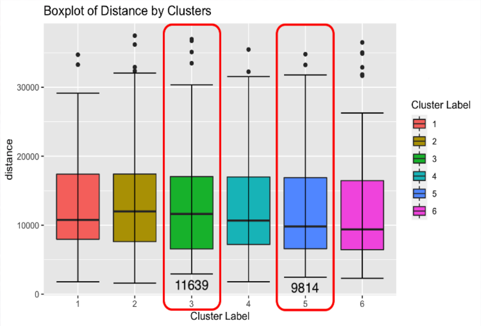
- Expensive vs Cheap
  - Modern vs Old
  - Hardcover vs Other
  - More than 5 years vs Less
  - High-rise vs Low-medium

- Farther from the epicenter…





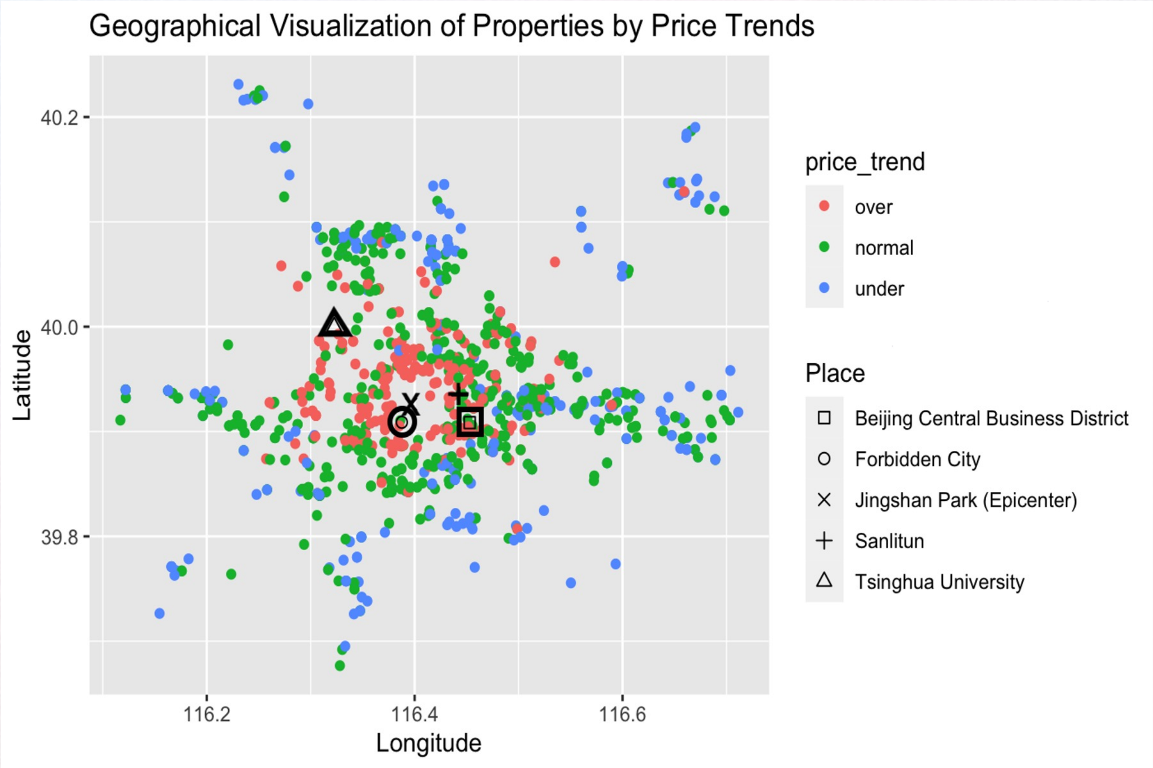




Price Trends

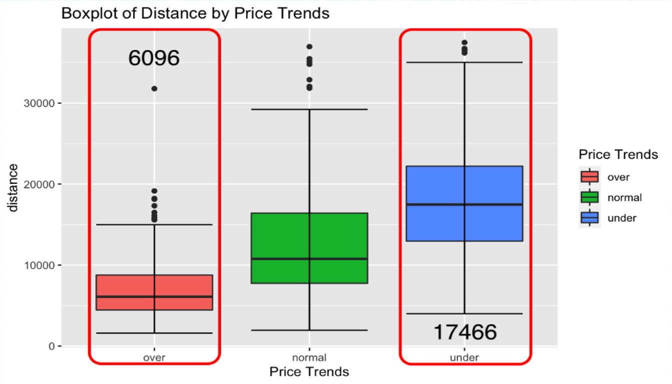
- Trends within a cluster
  - Price less than 25th percentile (Q1) → Under
  - Price greater than 75th percentile (Q3) → Over
  - Price in between the 25th and 75th percentile → Normal

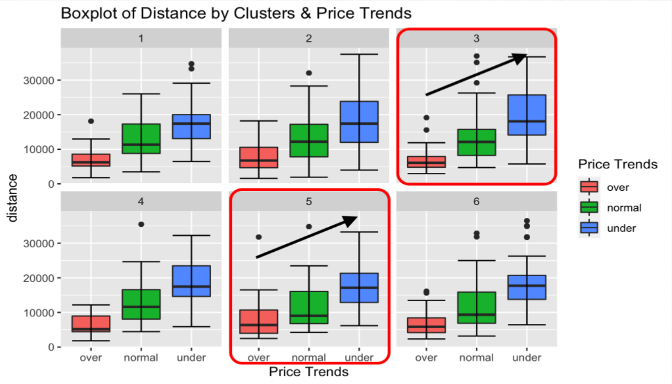
- Notable 'Price Trend' based on geographic

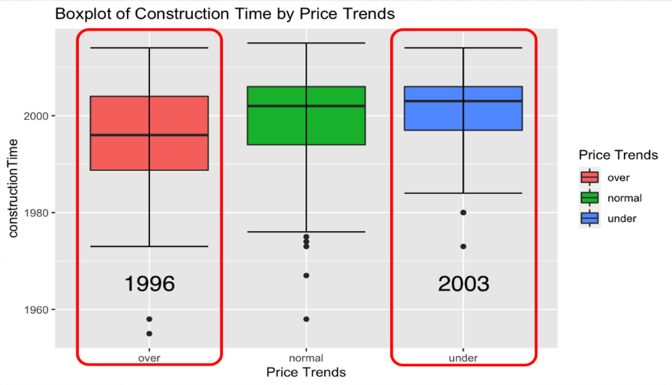


Price Trends vs X

- Distance aligning with Price Trends
  - Closer → Over
- Opposite trends
  - Over tends to be…
    - Old
    - Low-Medium stories
- City Planning Viewpoint
  - Center → Outside (Suburb)
  - Old & Low → Modern & High
  - More infrastructures around the epicenter







Takeaways

- Clusters
  - Possible Key factors
    - Construction Time / Renovation Conditions / Five Years Property / Floor #
  - Expensive vs Cheap
    - Modern vs Old / Hardcover vs Other (Rough) / More than 5 years vs Less / High vs Low
- Price Trends
  - Distance
  - Opposite factors : Construction Time / Floor #
  - City Planning Perspective
- Caveats
  - Data from 2011 to 2017
  - May not reflect current real estate market (e.g. Policy & Regulation)