Stat 432 Final Project

Beijing Housing Prices

Junseok Yang [jyang247], Himanshu Kumar [hk45], Matthew Jewell [mjewell3]

Project Description

Our team elected to work with a large dataset of real estate prices in Beijing. Real estate in a rapidly-growing country like China is an exploding industry, with the value of completed real estate in the country more than quadrupling in the ten years between 2004 and 2014, and continuing to grow.[1] In Beijing specifically, with a population of over 21 million, housing is in constant demand. This made Beijing an excellent subject for our project, and meant that effective tools for navigating the real estate market in the city could be useful for a great number of people.

Beginning with nearly 320,000 observations and 26 parameters, we aimed to clean the data and identify the most important parameters to make analysis computationally feasible while preserving the utility of the data and the algorithms we were creating. We had two objectives: one of which would rely on regression analysis, and another which relied on clustering algorithms. Our first goal was to create a regression model which could accurately predict the total cost of a real-estate listing, given parameters about the listing. The purpose of such a model would be twofold. First, someone planning to list a property could use the details of their intended listing to be given a reasonable price suggestion. Not only could this help landlords and tenants looking to sublease, but a real-estate developer could also use such an algorithm to predict their future income from rent if they undertake a construction project or acquisition. Alternatively, someone looking for housing could use such an algorithm to find what price they should expect, given their desired housing parameters, and they could adjust parameters to find what sort of listings they can expect to find within their budget.

The second goal for our project was to apply machine learning to meaningfully cluster existing real estate listings on the market. With those clusters created, our goal was to identify overpriced and underpriced listings, as compared to listings with similar parameters. With consistently updated data, investors could use this tool to identify properties that could be bought and flipped for a profit, and individuals could find more economical deals which would allow them to get the best quality housing for their budget.

Literature Review

To help understand what work others had already done with the Beijing housing dataset and to get ideas of what might or might not work well, we read through the notes of other programmers who had also attempted to predict housing prices given the same data. Some similarities appeared across the approaches we found, such as the use of gradient boosting regressors. Everyone also had to address the large amount of missing data, specifically for the "DOM" (Days on Market) variable, which helped our group reach a decision on how to handle that same complication and better understand where the missing data stemmed from.

[1] https://www.statista.com/statistics/243192/value-of-completed-residential-real-estate-in-china/

The first approach we examined found external data of second-hand home transactions from the National Bureau of Statistics.[2] They used this data to create a "growthRate" variable, calculated from monthly price growth in the period 2010-2017 and reported as relative to March 2010. The use of external data, while potentially effective, seems to go against the spirit of what we've been asked to do, and so our group noted this decision as potentially very effective but not something we will opt for in our own analysis. Instead, we focused on how they addressed missingness in the "DOM" variable, and how they utilized city landmarks. The authors addressed the missing "DOM" values by finding reported "DOM" values for observations that shared "district", "tradeYear", and "followers" values with missing "DOM" observations, and then assuming those "DOM" values would be equal.

The authors utilized the latitude and longitude data by plotting it and including four major locations (Beijing International Trade Center, Tiananmen Square, Temple of Heaven, and Xin Jiekou). They then found the Euclidean distance of the observations from each of those four locations. The distance from those locations did not prove to be a strong predictor of price, but our group planned to use longitude and latitude differently, so that wasn't concerning to us. In the end, the group's XGBoost (Extreme Gradient Boosting) model performed best.

The next group we examined also obtained their lowest RMSE using Extreme Gradient Boosting.[3] Their top three models all used boosting methods, but even their XGBoost model couldn't compete with the first group's, which may be a result of the first group's use of external data. With that said, the second group used Seaborn Relplot to group points by latitude and longitude into districts. This is more aligned to what our group is considering, but still not exactly what we are envisioning for how we will use latitude and longitude data. Regardless, it is an interesting use of the data and something our group will keep in mind when building our own model.

This group filled in "DOM" values with the mean values of groupings they made based on "district", "renovationCondition", and "buildingType" variables. What's more interesting is that this group seemed to have an explanation for all of the missing "DOM" values. They noted: "the DOM values were made to be calculated automatically with the listing date and the sold date values. Since the dataset is already a few years old, it's likely that lianjia's website structure changed in the meantime and made these DOM calculations impossible". My group had speculated that missing "DOM" values were perhaps a result of housing that was on the market for under a day, and had even considered filling in the missing observations with zeroes. If the second group's suggestion is correct, my group would have greatly diminished the accuracy of our data by filling in the null values with zeros.

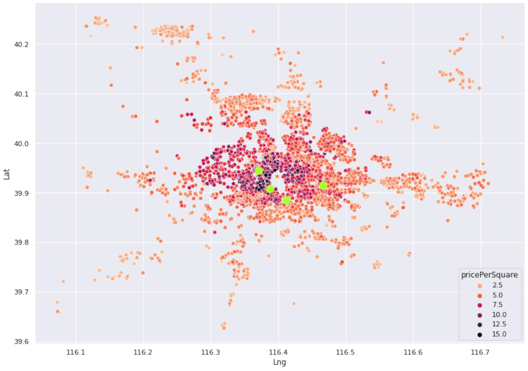
While other groups' work was reviewed, there is only one more worth mentioning, even though they did not get far enough to report an RMSE.[4] This third group took a unique approach to utilizing latitude and longitude data, which is still distinct from what my group had planned,

[2] https://www.kaggle.com/code/ericliu123/beijing-houseprice-predict

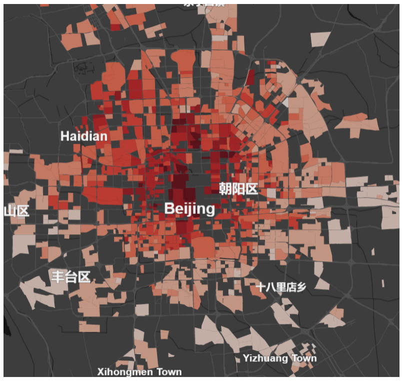
[3] https://www.kaggle.com/code/wutangfza/beijing-house-pricing-regression

[4] https://www.kaggle.com/code/alshan/beijing-housing-prices-on-a-map-with-spatial-join/notebook

but very interesting to note. The group spatially joined the Beijing real estate dataset with polygons form from Beijing city blocks. They were then able to calculate an average property price (whose units were simply referred to as “per square”) within any given city block and create a choropleth map of the city. The group’s resulting graphic would be excellent for easy, interactive use, and it could offer insights for my group to consider while creating our models. For example, it would seem that there is an epicenter of high property value, and values tend to decrease as properties are farther from that place. For this graphic alone, the third group’s work is worth considering while creating a model for Beijing housing prices.



Visualization of data with four Beijing landmarks (green dots).



Average price per square meter mapped onto Beijing street map.

Data Cleaning and Processing

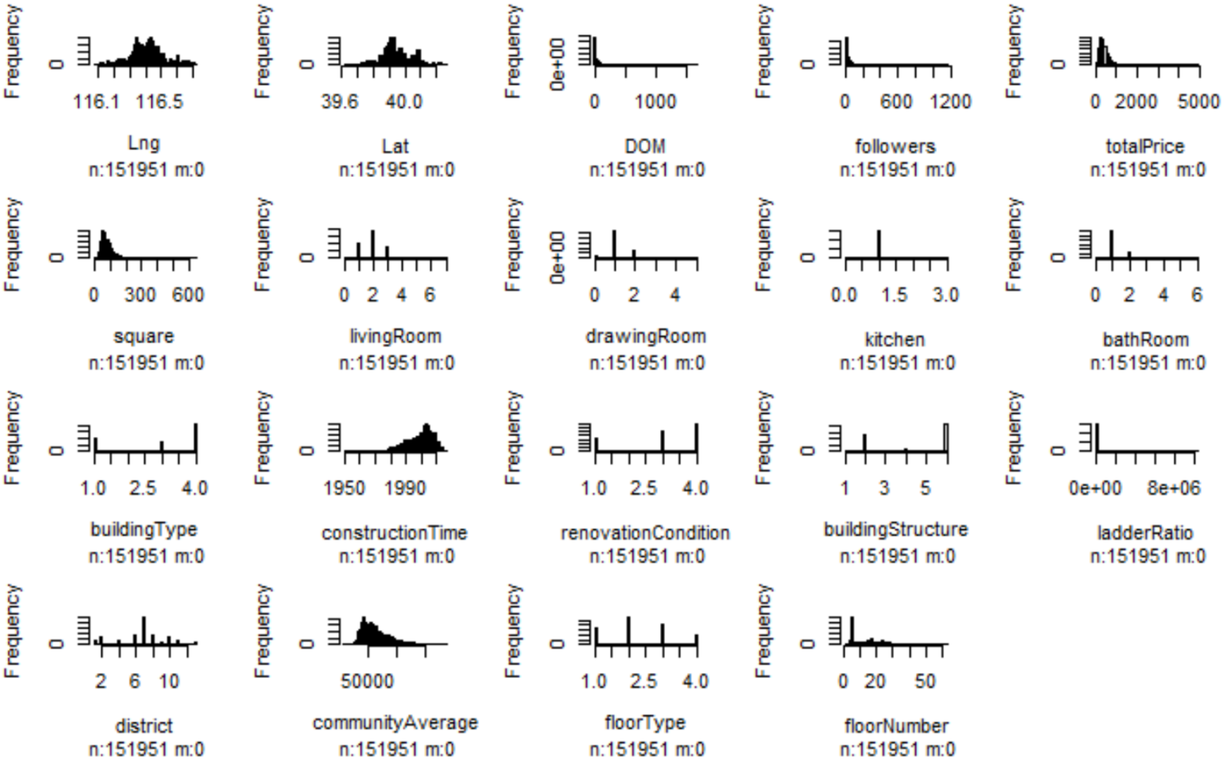
Our initial dataset had multiple variables which either needed to be refined, were unclear regarding what exactly they measured, or both. This was in part caused by data being drawn from another country with a language we were unfamiliar with, and in part due to confusing values observed for certain variables. The most troublesome example of translation difficulties stemmed from the “Floor” variable.

The “Floor” variable caused the most confusion for two reasons. First, while the observations did have numbers included in them, we didn’t know whether they referred to the number of floors in the structure, the number of floors in a listed unit, or what floor a listing was located on. Our uncertainty was compounded by the presence of Chinese characters which weren’t properly read by R, becoming unidentifiable, and a lack of notes regarding the “Floor” variable in the dataset’s description. However, during our literature review, we discovered that the Chinese characters had four meanings: bottom, low, middle, or high. Each of these four floor types were written as Chinese characters and combined with a number as a string. We split these strings into two new variables: “floorType” and “floor”Number”.

Another sizable issue with our dataset was missingness. About 50% of the observations in our dataset had missing values for at least one variable, with the largest culprit being “DOM” (Days on Market). “DOM” was missing in 49.54% of observations, with “buildingType” - the variable with the next most missing values - only missing about 0.63%. During our literature review, we found that “DOM” was an important variable, and a strong estimator of the price we were trying to predict, so we were reluctant to get rid of it entirely. We reasoned that, with around 160,000 observations still having “DOM” values, we might get better results dropping all observations whose “DOM” values were missing, rather than using imputation or dropping the variable.

Still, to be safe, we decided to try different datasets either imputing “DOM”, dropping observations with missing values, or dropping the entire variable. The decision was made for us when computational limitations kept us from analyzing the data with all 320,000 observations and “DOM” included, as R would crash whenever we tried. We had already done various tests such as checking variance inflation factors, AIC, and Mallows’s Cp, and we found ourselves in agreement with the literature we reviewed in finding that “DOM” was too important to remove entirely. Instead, we moved forward using only 160,000 observations, which alleviated the computational burden on our computers.

One final piece of our preprocessing was addressing latitude and longitude, and converting them into a variable we wanted to use. Initially, there was just concern regarding any linear use of latitude and longitude, as moving too far in any direction would put us outside the city, so we were concerned about a linear coefficient for either variable. However, in our literature review, we found a group that identified four important landmarks around Beijing (Beijing International Trade Center, Tiananmen Square, Temple of Heaven, and Xin Jiekou) and incorporated the distance from them into their analysis. Another group had imported an actual map of the city, using the real streets to separate the different blocks and make a heat map of prices. That map seemed to have a clear epicenter in a certain part of the city around which prices seemed to be centered, with the highest prices being the closest listings. So, combining the ideas of those other groups, we used Google Maps to find what landmark was located where all the listing prices radiated from, and we found Jingshan Park. To account for the curvature of the Earth and its non-spherical shape, we used the geosphere package and the park’s latitude and longitude to generate a new “Distance” variable for every observation, recording how far each listing is from Jingshan Park.

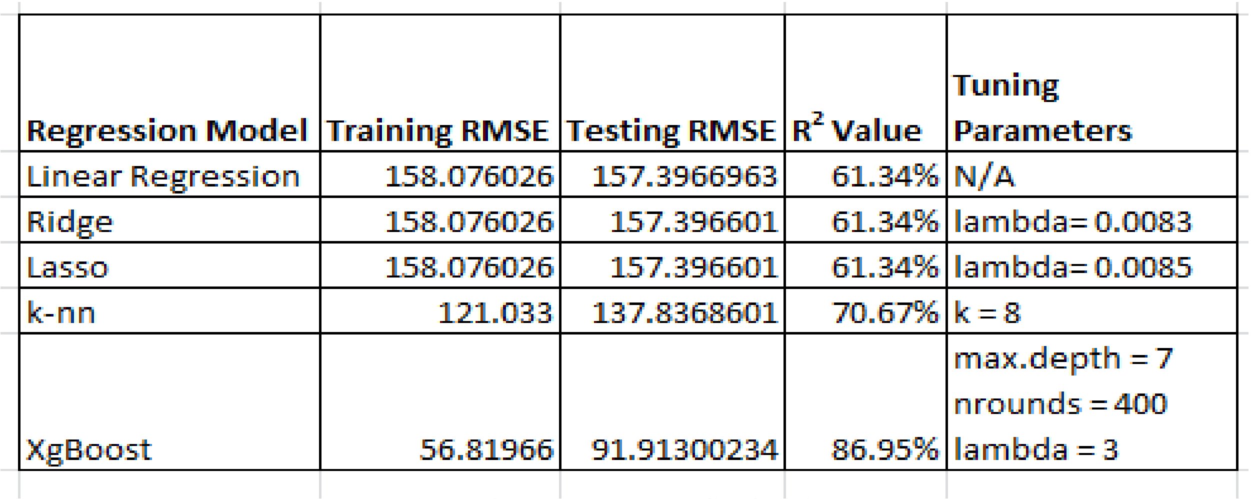


A visualization of our data, with incomplete observations excluded, as part of our initial exploration of the dataset.

Findings and Discussion

In our regression analysis, we performed K-Nearest Neighbor (KNN), Extreme Gradient Boosting (XGB), and a Penalized Linear Regression to build models predicting total prices of housing listings. We identified our top ten most powerful parameters for predicting our regression outcome using AIC and Mallows's CP. We found these parameters to be: "DOM", "livingRoom", "drawingRoom", "bathRoom", "constructionTime", "renovationCondition", "fiveyearsProperty", "communityAverage", "floorNumber", and "distance"–the variable which we created to measure distance from Jingshan Park. We removed "ladderRatio" based on our AIC results, and noted that "Lng" (longitude) would also ideally be removed. The latter was a non-issue, as we had already planned to replace both latitude and longitude with our "distance" variable, which, as noted above, was determined to be one of our most useful parameters.

We were unsurprised to find that our Extreme Gradient Boosting model performed the best, after seeing similar results in our literature review. Our best performing model was able to achieve an RMSE of 91.833, which was better than the typical RMSE we found from other groups. With total price values measured by the thousands, we were satisfied with this level of accuracy, although we will address in our conclusion below how we might be able to improve our 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

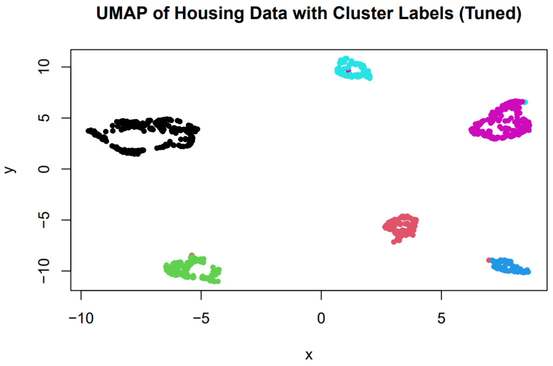
Regression results report. Extreme gradient boosting performed best of the approaches we attempted.

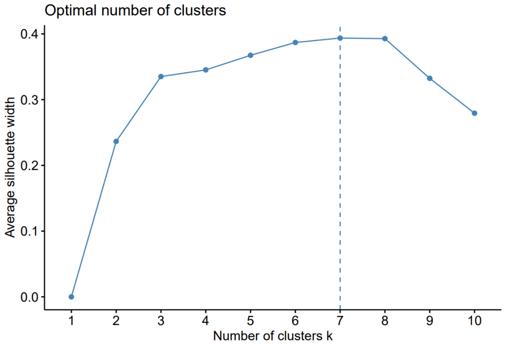
Given the dataset we were using was large and complex with many parameters, we ran into issues with computational costs and limitations. We attempted to fit our raw data to K-Means and hierarchical clustering naively, but each attempt resulted in R aborting its session. Fortunately, we had identified our ten most important variables mentioned above during the regression portion of our analysis, so we used those variables for clustering. We also opted for K-means clustering over hierarchical clustering, and we used simple random sampling from our larger dataset–both to manage our computational limitations.

Due to K-means clustering using Euclidean distance, all variables should be numeric. This added an extra data processing step, as some variables were technically numeric, but wouldn't make sense in an arithmetic analysis context. For example, "fiveyearsProperty" used 0 and 1 to represent whether a property had been owned for 5 years or not. Additionally, K-means clustering could be skewed by a variable with a large scale compared to the other parameters, so we needed to scale our variables to proceed.

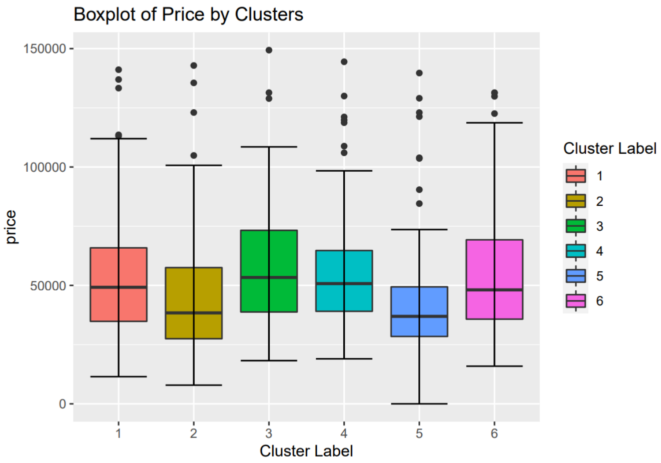
To make our data suitable for clustering, we considered Gower's distance and one-hot encoding. One-hot encoding converts factor variables into several dummy variables, but the curse of dimensionality can quickly become problematic as a column is expanded based on how many levels its factor variable has. Gower's distance, which calculates the dissimilarity between non-numeric variables, allows for their conversion to be numeric. In the process, some information is lost, and Gower's distance won't reflect nuanced similarities between observations. After weighing our options, especially in the context of computational limitations we had run into, we proceeded with Gower's distance.

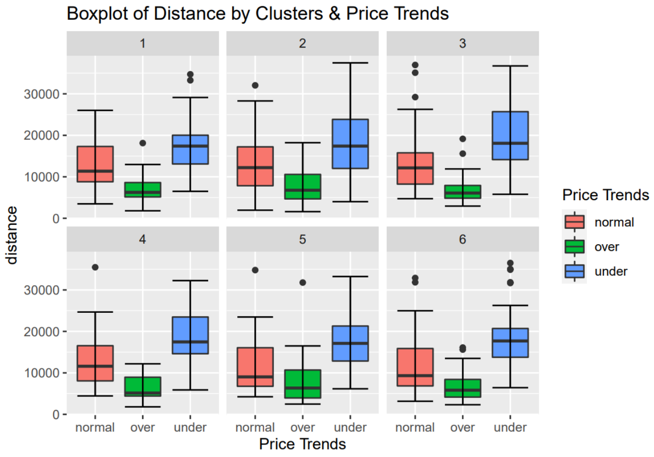
We visualized different numbers of clusters, using the "Elbow" method, a Silhouette Score, and UMAP. These methods suggested similar optimal numbers of clusters (five to seven), and we settled on six. We visualized the summary statistics by cluster and have included the price boxplots below. Given the purpose of our clustering (to identify overpriced and underpriced listings), the clusters did not necessarily need to have very price distributions. However, as one might expect listings with similar attributes to be priced similarly, it is somewhat surprising that the prices captured by each cluster are so similar, and so widely spread. This may be a result of our computational limitations only allowing us to use smaller samples of around one thousand observations.

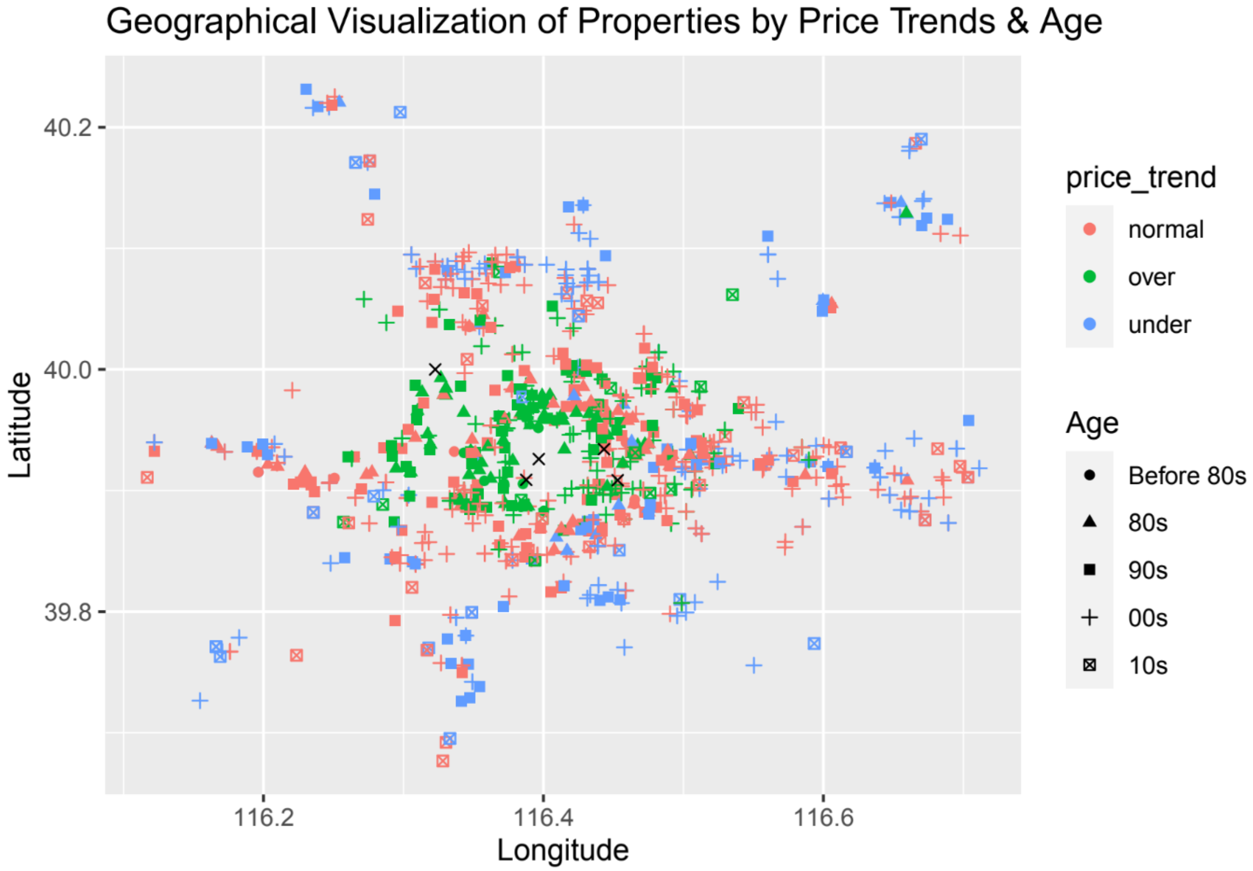




Within each cluster, we separated the observations into price trends: "normal", "over", and "under". Summary statistics were visualized with these groupings as well. One interesting observation was that overpriced listings had the smallest "distance" variables in every cluster, normally priced listings were supporting our belief that distance from Jingshan Park would be a strong predictor of housing price. We also noted that the overpriced listings in each cluster tended to be older than normally priced and underpriced listings, and that overpriced listings had relatively lower "floorNumber" values. We found these trends to be surprising, as we had expected that newer properties would be more modern and thus more expensive. There could be a relationship between the value of being near the city's center (i.e. Jingshan Park) and the natural progression of building the city outward leading to older listings being at the heart of the city.







Many underpriced listings were more modern, and many overpriced listings were older.

Conclusion and Future Next Steps

Our team has created two tools for both commercial and consumer use, increasing transparency for both landlords and tenants. Our regression allows renters to accurately see how much an apartment with their desired parameters should cost, which can also alert renters if their current payments are too high (or too low). For real estate developers, the regression model provides a tool with which they can calculate anticipated future cash flows if they purchase a residential property or undertake a construction project.

Similarly, our clustering model as a tool will allow developers and renters to identify properties that are currently listed at underpriced levels. For renters, this presents an opportunity for acting on a deal and maximizing how far their budget goes. For landlords looking to expand, this tool will highlight properties that they can buy for less than what the properties are expected to cost, effectively presenting investment opportunities.

The limitations of our models largely stemmed from missingness and computational resource shortcomings. To improve on the work we've done, our next steps should be trying to fill in missing data and obtaining resources that can process larger amounts of information as well as more intensive processes such as hierarchical clustering, which we were unable to use due to a lack of computational resources.

The models we have created also rely on having constantly updated information to keep track of current market trends and identify underpriced/overpriced listings. To the extent that our intentions were to create tools with practical uses, our models will lose their worth without these updates. The best course of action would perhaps be partnering with a real estate app similar to Zillow or Redfin, to make use of their up-to-date databases while providing their apps additional functionalities. Such a partnership would also likely provide the computational resources that we were lacking.