Predicting Rental Prices in New York City

Metis

Heeyewon Jeong

Design

Backstory

Apartment rental prices have changed amid the pandemic.

A real estate company wants to inform post-pandemic rental prices of New York City apartments for new potential tenants and landlords.

Project Goal

To see how the rental prices in New York City can be predicted by apartment conditions and locations.

Data Collection

Data

1500+ apartments from CityRealty.com

Tools

- Beautiful Soup and Selenium
- Scikit-learn and Statsmodels
- Matplotlib and Seaborn
- NumPy and Pandas

Target

Monthly rental price

Features

- Number of bathrooms
- Number of bedrooms
- Square feet
- Number of available amenities
- Year of building built
- The number of units in building
- Floor level of building
- Distance to the closest stations
- Neighborhood
- Type of fee

  (i.e., no broker fee, co-op fee, or condo fee)

Data Cleaning and EDA

Drop missing data

The website did not consistently list the same information for all apartments

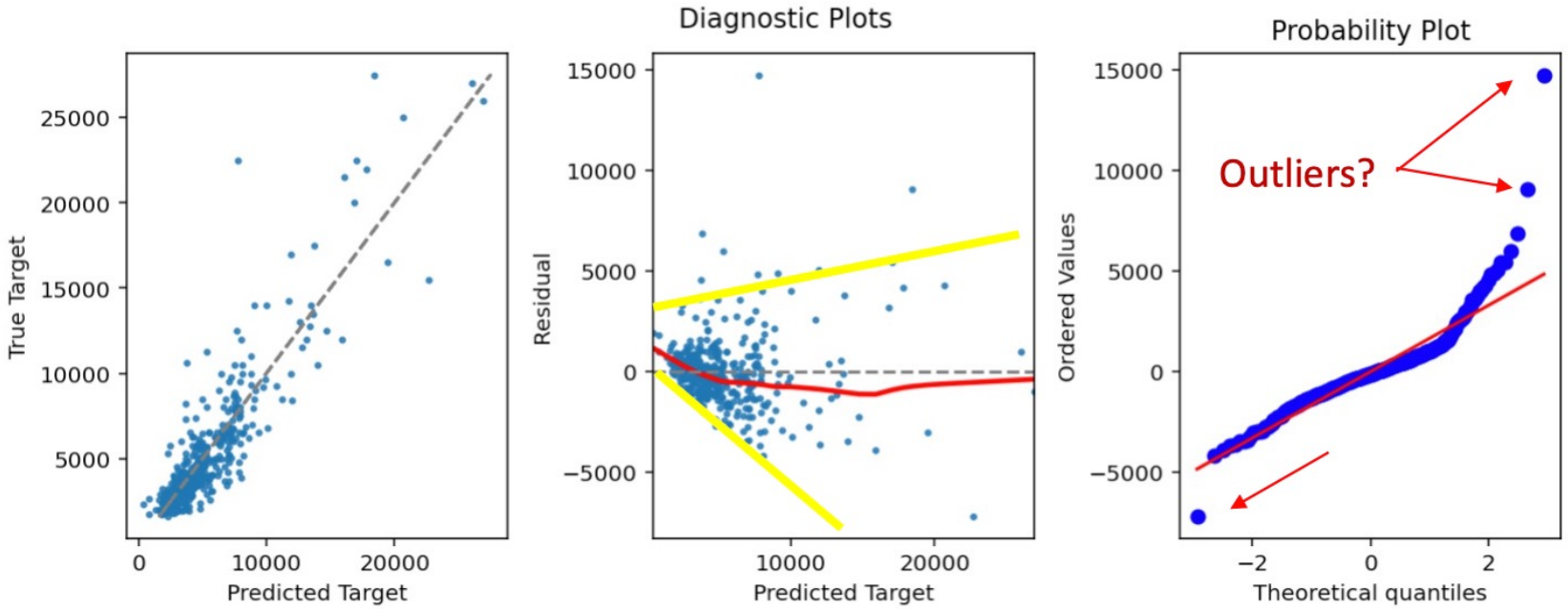
Number of Bedrooms as a categorical variable

More than 6 bedrooms were not counted on the website and listed as ‘6+’

Grouping neighborhoods

A total of 60 Number of Neighborhood were grouped by borough for simpler interpretation

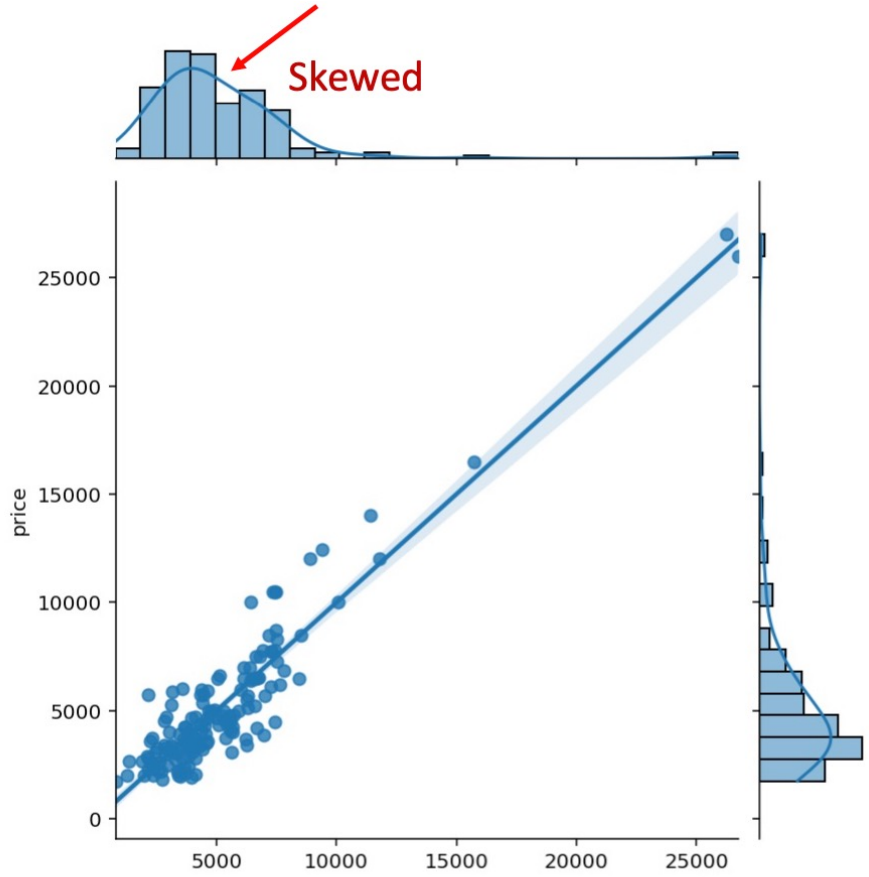
Baseline & Diagnostic Plot



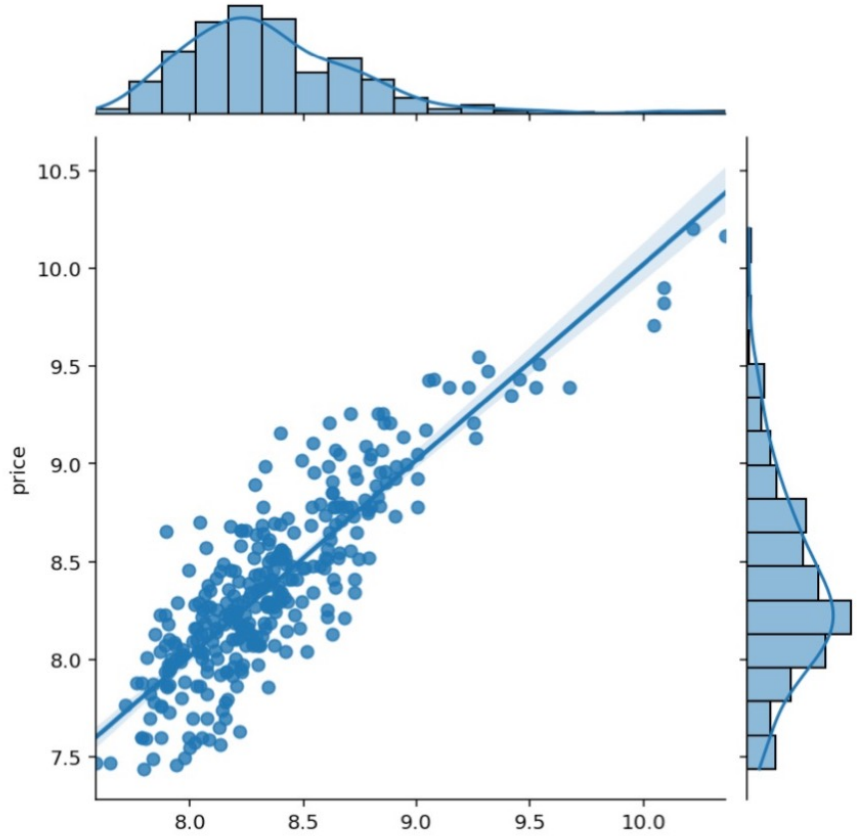
- ✓ Due to the outliers, only the apartment **less than $15,000** were considered.
  Caveat : Due to the outliers dropped, the current analysis was limited to predict apartments with bedrooms less than 5.

- ✓ To predict better, insignificant variables were dropped.
  I.e., number of bedrooms and units, types of fee, and floor level of apartment

Ordinary Least Square



Log Transformation



OLS
Train: 0.77
Val: 0.702

RMSE: 1151.743
MAE: 853.685

OLS with a log transformation
Train: 0.723
Val: 0.702
Test: 0.702

RMSE: 0.237
MAE: 0.181

Best Model

Best predictive Model: Simple vs. Ridge Regression with K-Fold

```
Simple mean cv r^2: 0.678 +- 0.093
Ridge mean cv r^2: 0.678 +- 0.093

Simple Regression test R^2: 0.723
Ridge Regression test R^2: 0.723
```

The Futures selected

- ✓ Number of bathrooms
- ✓ Square feet
- ✓ The year of building built
- ✓ Number of amenities
- ✓ Closest distance to nearby stations
- ✓ Location (i.e., Boroughs)

Lasso Regression for the best interpret model

```
Lasso model alpha: 0.01
MAE: 0.178
r^2: 0.704
```

The Futures selected

- ✓ Number of bathrooms
- ✓ Square feet
- ✓ The year of building built
- ✓ Number of amenities
- ✓ Closest distance to nearby stations

The ***least*** complex model with the zero coefficient for 'Manhattan'

Conclusions

Takeaways

- The following were not significantly associated with the rental prices.
  - Types of a fee
  (e.g., no fee versus broker fee)
  - The total number of units in an apartment
  - Taller apartments are not necessarily more expensive.

- Rental prices in Manhattan were more difficult to predict than the other boroughs
  - Higher density with diverse population brings higher variance?

Future work

- Adding socioeconomic data per zip code
  - Median income
  - Age
  - Education Level

- The distance of closest stations was negatively associated with rental price and did not necessarily increase the price.
  - Number of train lines or stations nearby?

- Accurately counted number of bedroom as a continuous variable