NYC Property Sale Price Prediction

Congyun Jin(cj2164)
Fan Zhang(fz2068)
Siyu Shen(ss14359)
Yi Wen (yw5280)
Zhuoyuan Xu (zx1137)

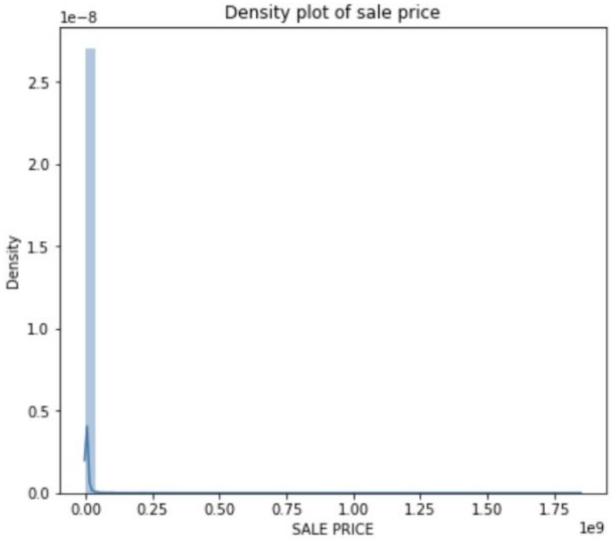
Data Understanding/Preprocessing

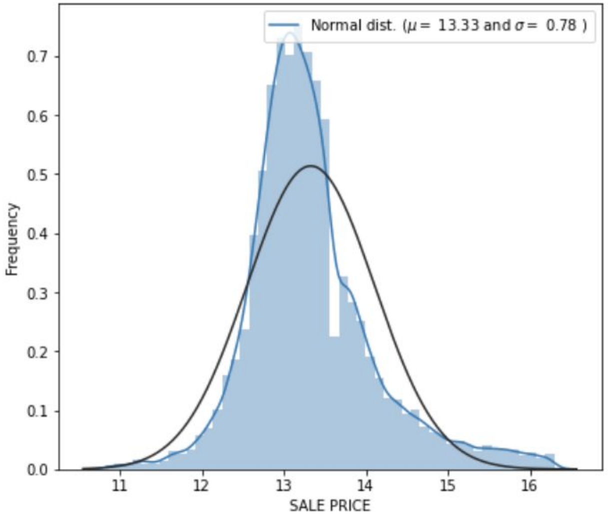
- The original data comes from the NYC Department of Finance
- It contains 167719 pieces of sales information of properties sold in New York City from January 2018 to December 2019
- After certain data cleaning processes, (drop missing values; duplicates; outliers) the cleaned data contains 81030 rows and 19 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167719 entries, 0 to 167718
Data columns (total 21 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   BOROUGH                         167719 non-null  int64
 1   NEIGHBORHOOD                    167719 non-null  object
 2   BUILDING CLASS CATEGORY         167719 non-null  object
 3   TAX CLASS AS OF FINAL ROLL 18/19  167439 non-null  object
 4   BLOCK                           167719 non-null  int64
 5   LOT                             167719 non-null  int64
 6   EASE-MENT                       0 non-null       float64
 7   BUILDING CLASS AS OF FINAL ROLL 18/19  167439 non-null  object
 8   ADDRESS                         167719 non-null  object
 9   APARTMENT NUMBER                36979 non-null   object
 10  ZIP CODE                        167710 non-null  float64
 11  RESIDENTIAL UNITS               155411 non-null  float64
 12  COMMERCIAL UNITS                155411 non-null  float64
 13  TOTAL UNITS                     155411 non-null  float64
 14  LAND SQUARE FEET                155410 non-null  float64
 15  GROSS SQUARE FEET               155411 non-null  float64
 16  YEAR BUILT                      162787 non-null  float64
 17  TAX CLASS AT TIME OF SALE       167719 non-null  int64
 18  BUILDING CLASS AT TIME OF SALE  167719 non-null  object
 19  SALE PRICE                      167719 non-null  object
 20  SALE DATE                       167719 non-null  object
dtypes: float64(8), int64(4), object(9)
memory usage: 26.9+ MB
```

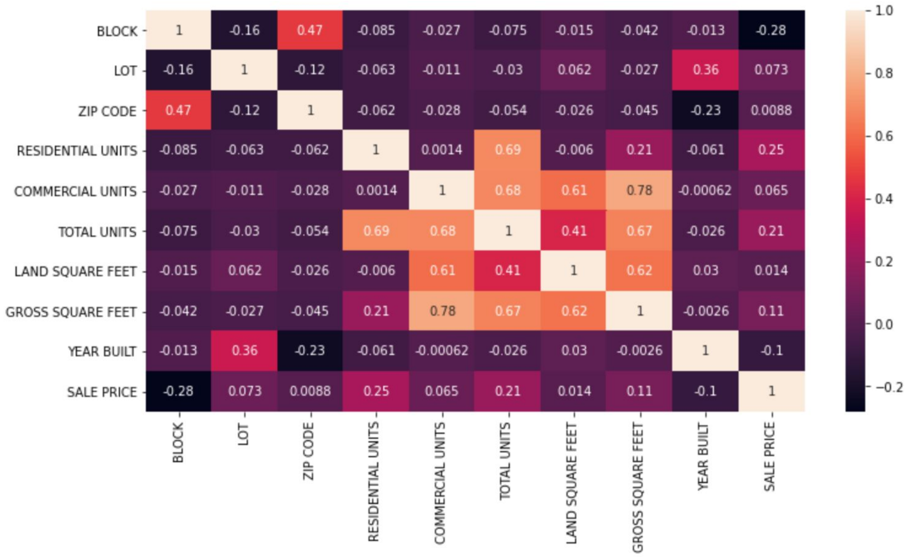
EDA: Property Sale Price (Target variable)

- The target variable is the sale price of each property.
- The distribution of raw data is really sparse.
- We set a reasonable range: \$50000 (38% rows dropped) ~ \$12M (0.8% rows dropped)
- Perform log transformation





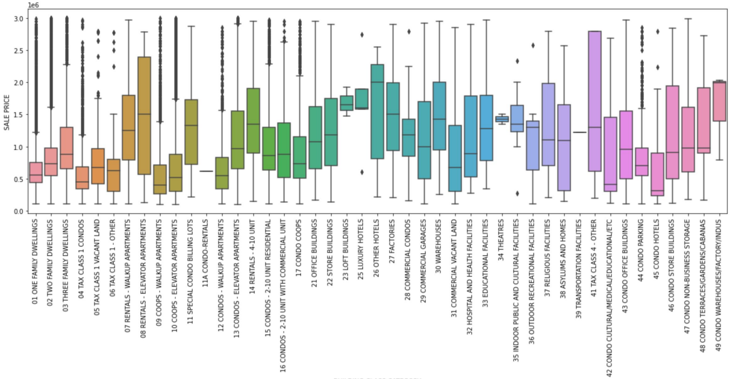
EDA: Correlational Analysis

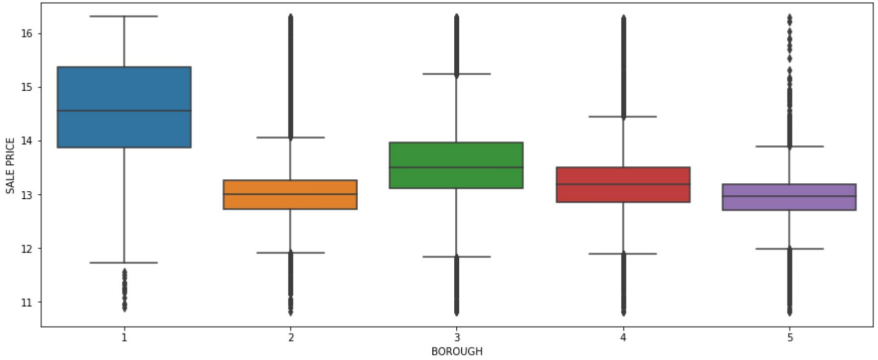


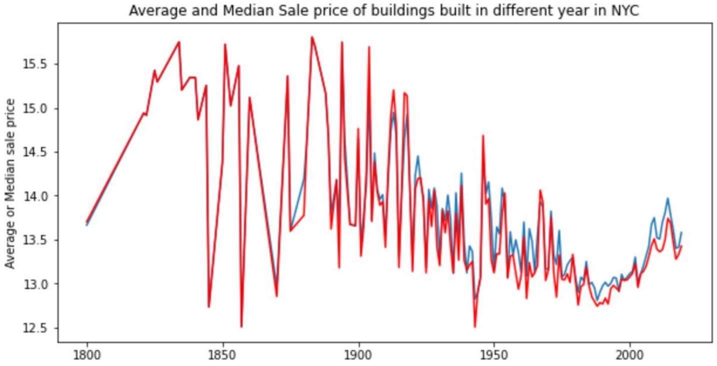
Before dropping features

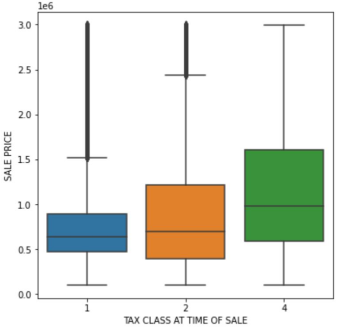
After dropping features

EDA: Predictive features





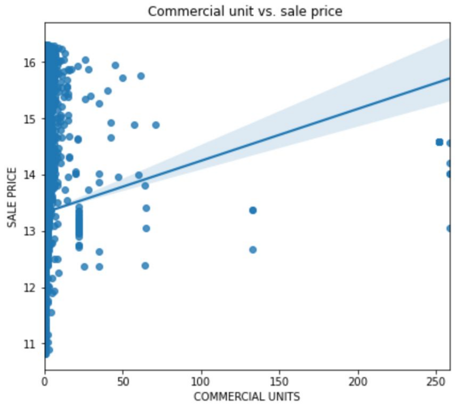


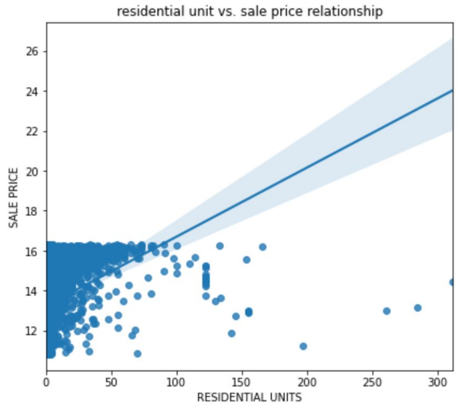


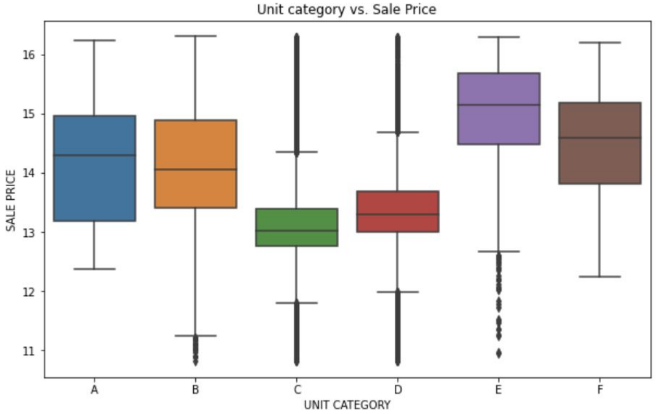
EDA: Commercial, Residential Units

- The pattern is opaque and there are lots of 0s, 1s in each plot. Hence we will classify them into six groups in *Feature Engineering process*

UNIT TYPE	CRITERIA
A	Commercial Units > 10
B	0< Commercial Units <= 10
C	Commercial Units = 0 and Residential Units= 1
D	Commercial Units = 0 and 1 < Residential Units < 10
E	Commercial Units = 0 and Residential Units >= 10
F	Commercial Units = 0 and Residential Units





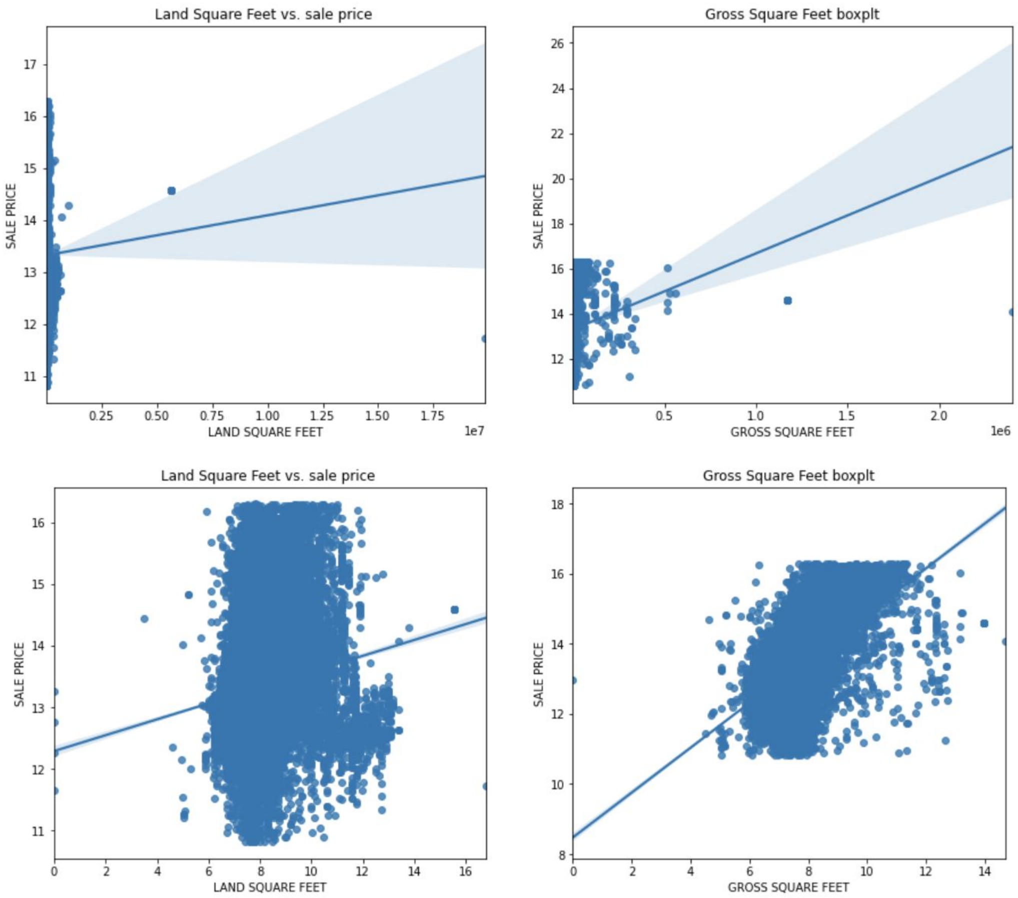


Feature Engineering

- Classify Residential Units and Commercial Units (mentioned before)

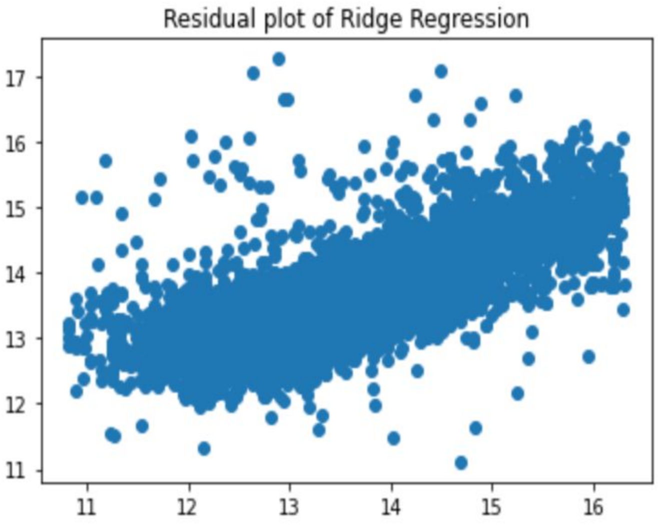
- One-hot coding on categorical variables

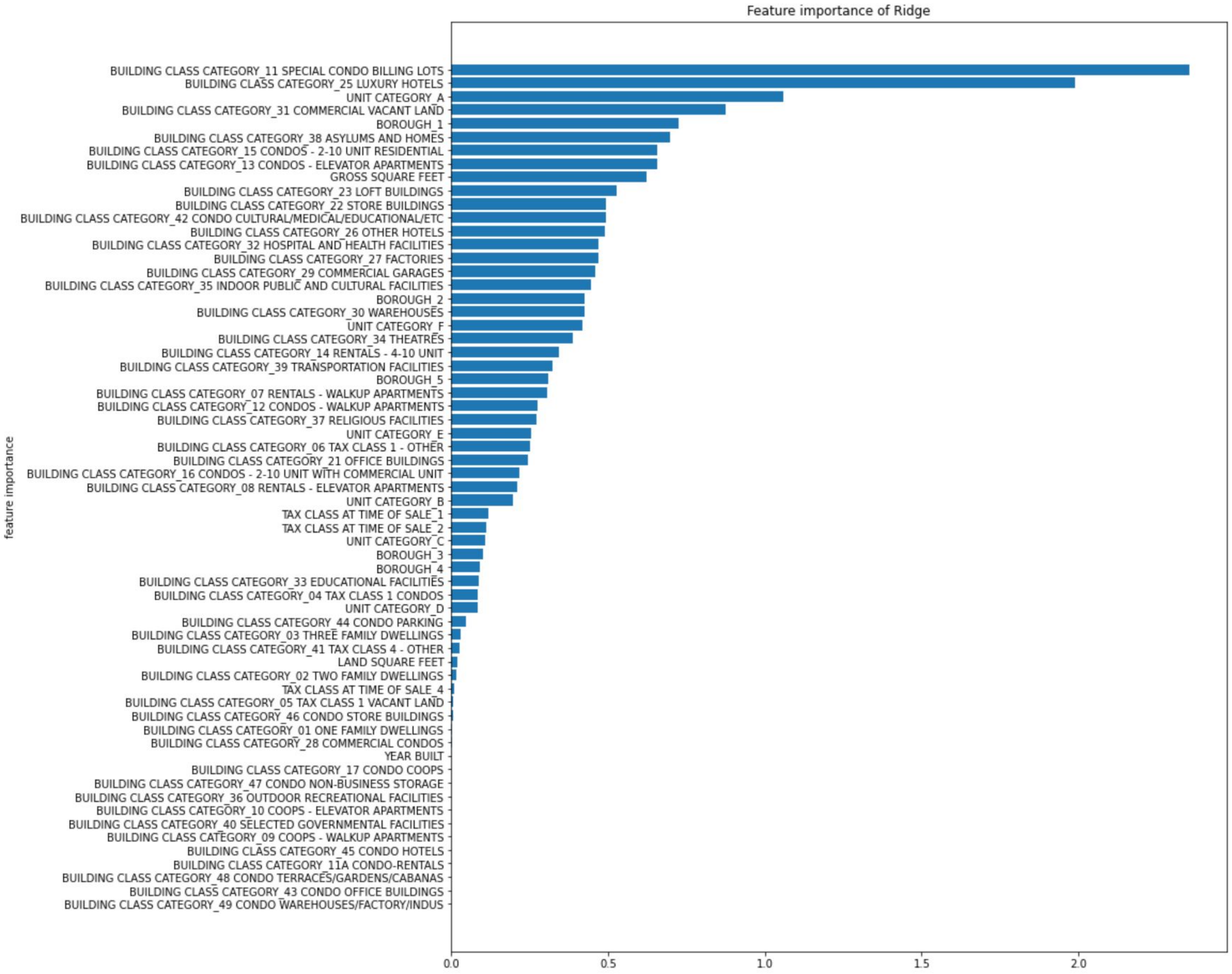
- Log transformation on numerical features (Land square feet, Gross square feet)



Model Performance

Model	MSE	R2
Linear	0.2840	0.5301
Lasso	0.4296	0.2894
Ridge	0.2840	0.5302
Robust	0.3088	0.4891
Elastic Net	0.4180	0.3086





Property Pricing Drivers: Sorted by Importance

Discussion

Models Assumptions Check:

- ➢ Normality: normal distribution of the target variable
- ➢ No Multicollinearity: heatmap
- ➢ Homoscedasticity: normal residual plots
- ➢ Independence: feature engineering on dependent variables

Future improvements:

- ➢ More precise spatial scales
- ➢ Add economic conditions indicator
- ➢ More reasonable binning and weighting

THANK YOU!

Q&A