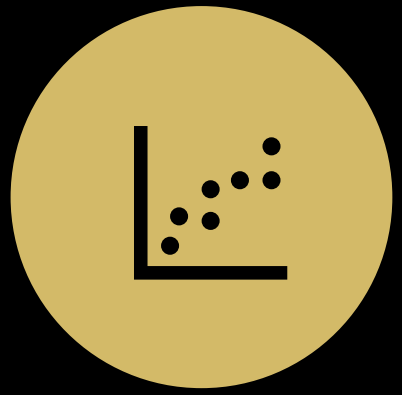
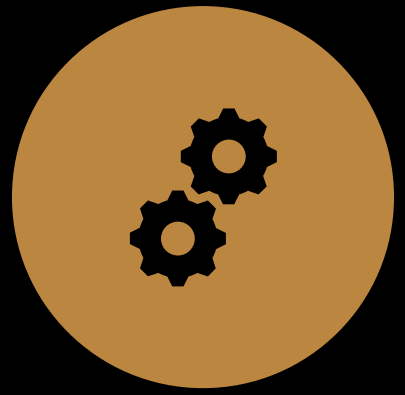
Staymore data analysis and machine learning

Team 7

SECTIONS



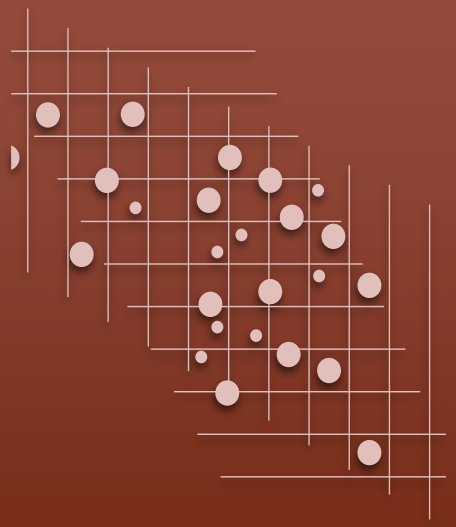
DATA EXPLORATION



FEATURES ENGINEERING

MODEL PREDICTION

DATA EXPLORATION



We chose to examine the price path in relation to the following features:

- Bedrooms
- Beds
- Bathrooms
- Bathrooms Type
- Longitude
- Latitude
- Property Type
- Room Type
- Number of Reviews
- Number of Reviews in 30 days
- Reviews Score (Rating, Accuracy, Check in, Cleanliness, Communication, Location, Value)
- Neighborhood
- Accommodates
- Availability in 365 days
- Reviews per Month
- Count of Host Listings
- Instant Bookable
- Amenity Count

THE DATASET

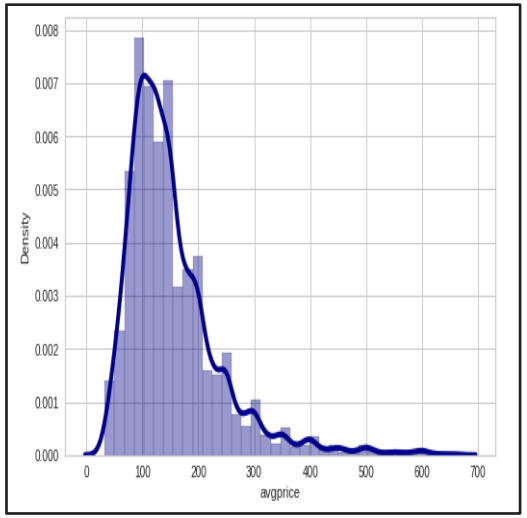
- At a first glance we can see the info of the features

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 18777 entries, 0 to 18781
Data columns (total 28 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Property_Key                    18777 non-null  int64
 1   Property_ID                     18777 non-null  int64
 2   Latitude                        18777 non-null  float64
 3   Longitude                       18777 non-null  float64
 4   Accommodates                    18777 non-null  int64
 5   Availability_365                18777 non-null  int64
 6   Has_availability                18777 non-null  int64
 7   Reviews_per_month               18777 non-null  float64
 8   Calculated_host_listings_count  18777 non-null  int64
 9   Property_type                   18777 non-null  object
 10  Instant_bookable                18777 non-null  int64
 11  Room_type                       18777 non-null  object
 12  Bedrooms                        18777 non-null  int64
 13  Beds                            18777 non-null  int64
 14  Bathrooms                       18777 non-null  float64
 15  Bathrooms_type                  18777 non-null  object
 16  Neighbourhood_cleansed          18777 non-null  object
 17  Number_of_reviews               18777 non-null  int64
 18  Number_of_reviews_l30d          18777 non-null  int64
 19  Review_scores_rating            18777 non-null  int64
 20  Review_scores_accuracy          18777 non-null  int64
 21  Review_scores_checkin           18777 non-null  int64
 22  Review_scores_cleanliness       18777 non-null  int64
 23  Review_scores_communication     18777 non-null  int64
 24  Review_scores_location          18777 non-null  int64
 25  Review_scores_value             18777 non-null  int64
 26  Amenity_Count                   18777 non-null  int64
 27  avgprice                        18777 non-null  float64
dtypes: float64(5), int64(19), object(4)
memory usage: 4.2+ MB
```

```
count    18777.000000
mean       162.896918
std        174.302646
min          5.000000
25%         99.032877
50%        134.332425
75%        189.000000
max       8000.000000
Name: avgprice, dtype: float64
```

→ Price Description

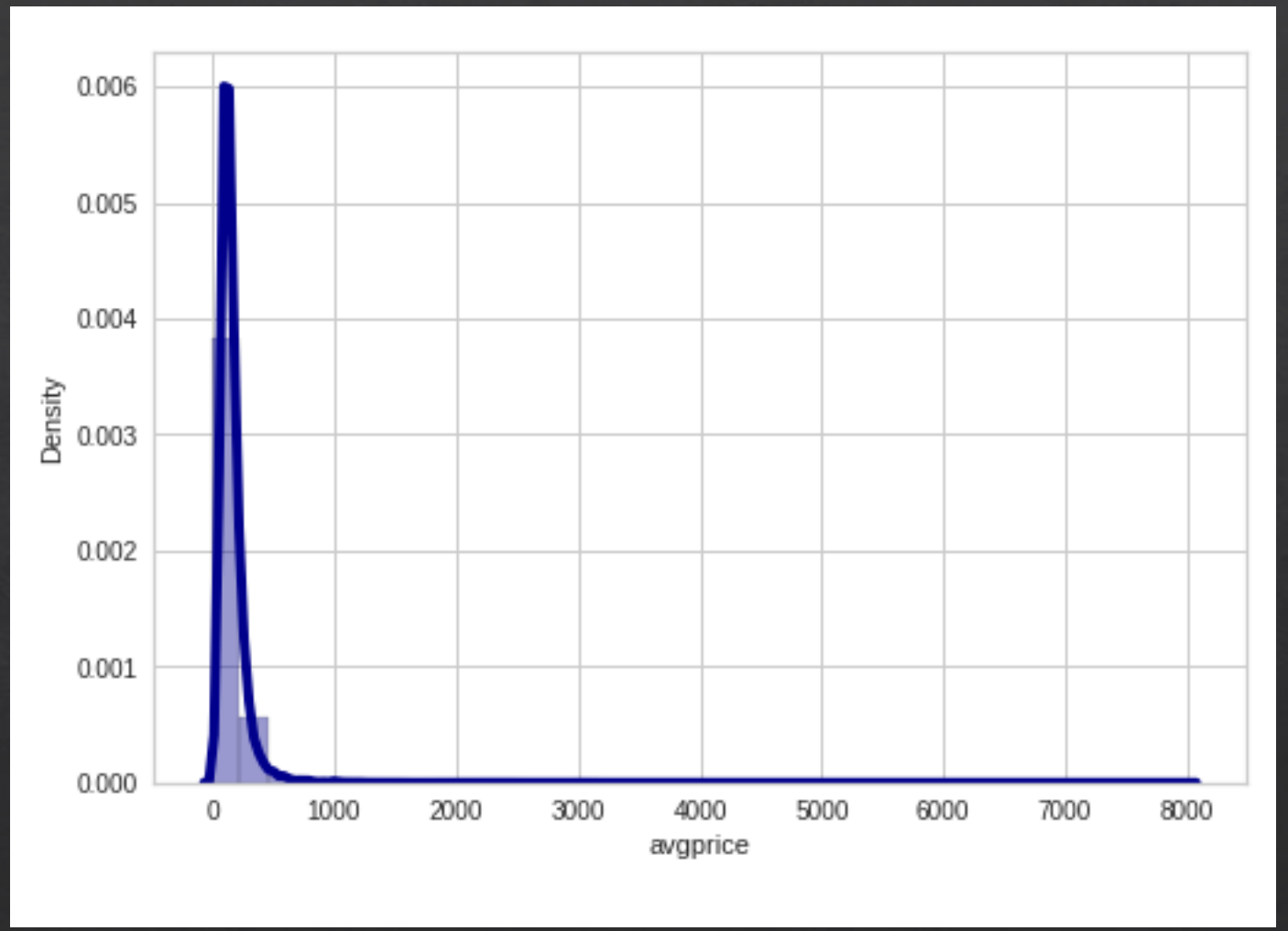
Price Distribution ←



20/12/2020

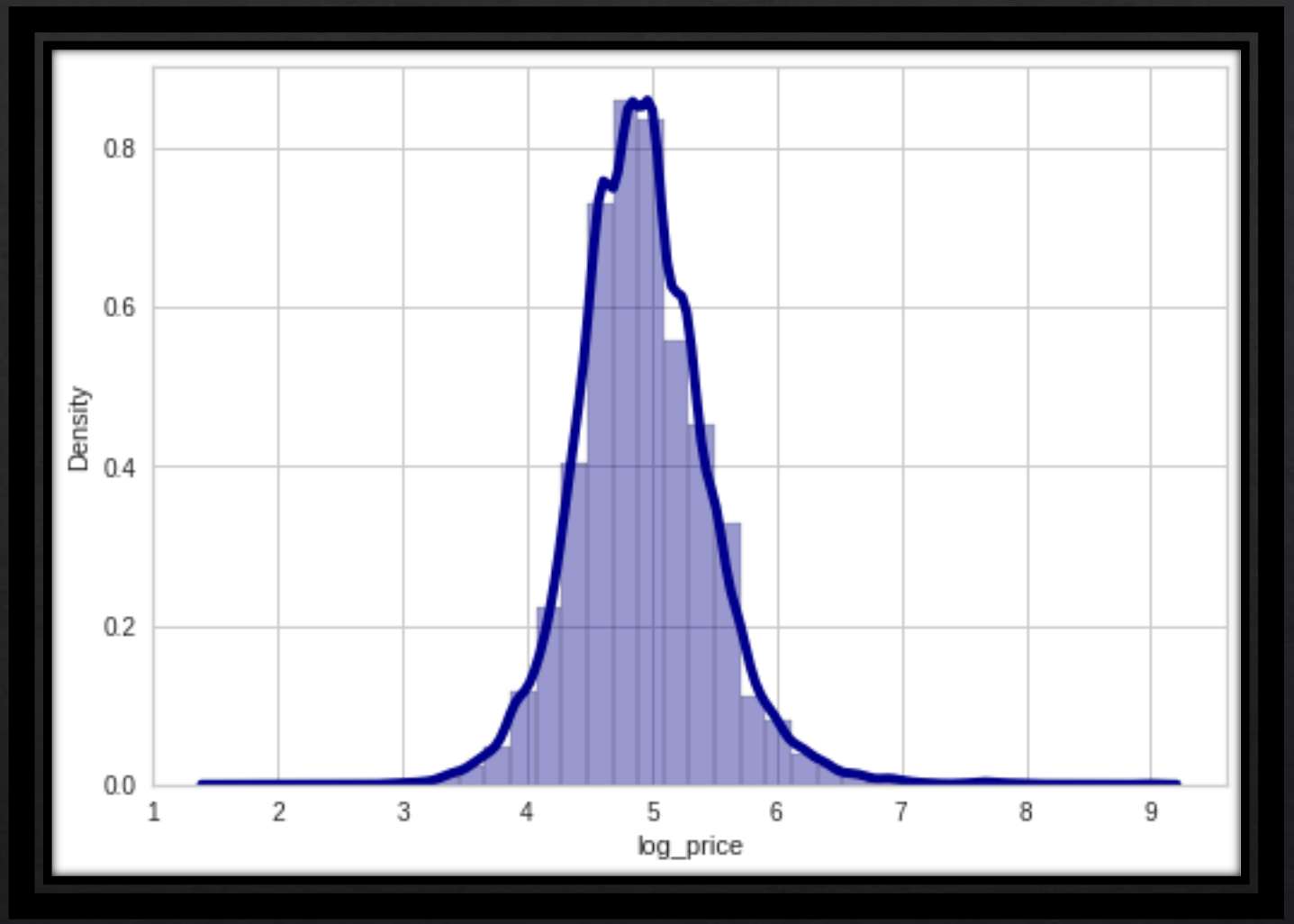
Price and log Price

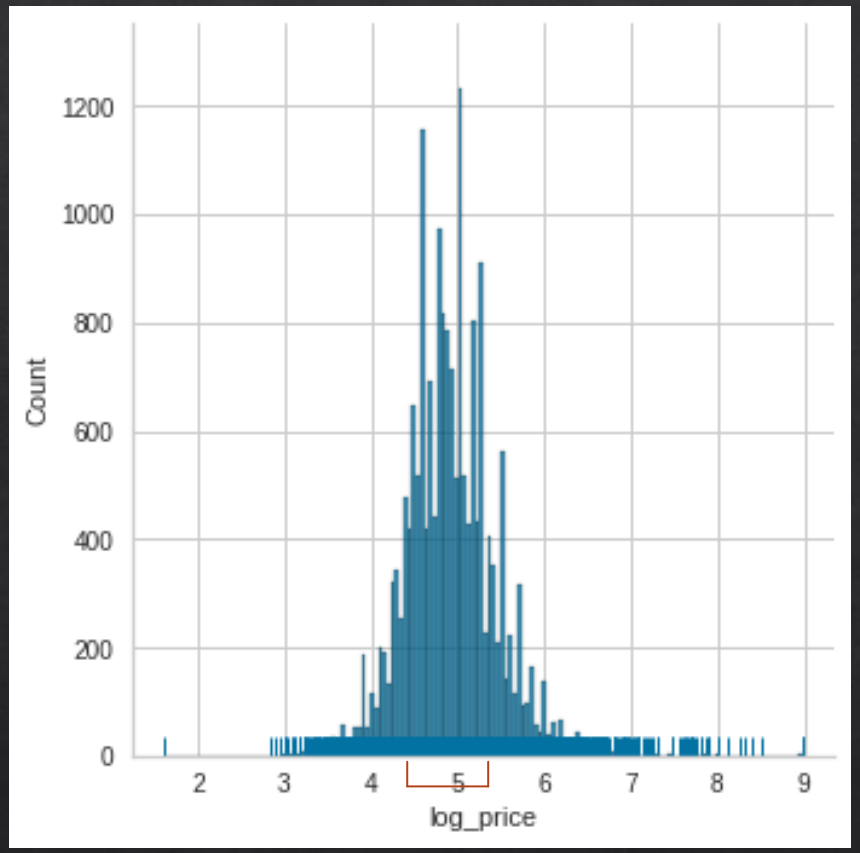
Description	Price
Mean	162.8969
Standard Deviation	174.3026
Range	7995
Median	134.3324
1st Quantile	99.0328
3rd Quantile	189



Price and log Price

Description	Price	Log Price
Mean	162.8969	4.924357
Standard Deviation	174.3026	0.530360
Range	7995	7.3777
Median	134.3324	4.900318
1st Quantile	99.0328	4.595452
3rd Quantile	189	5.241747



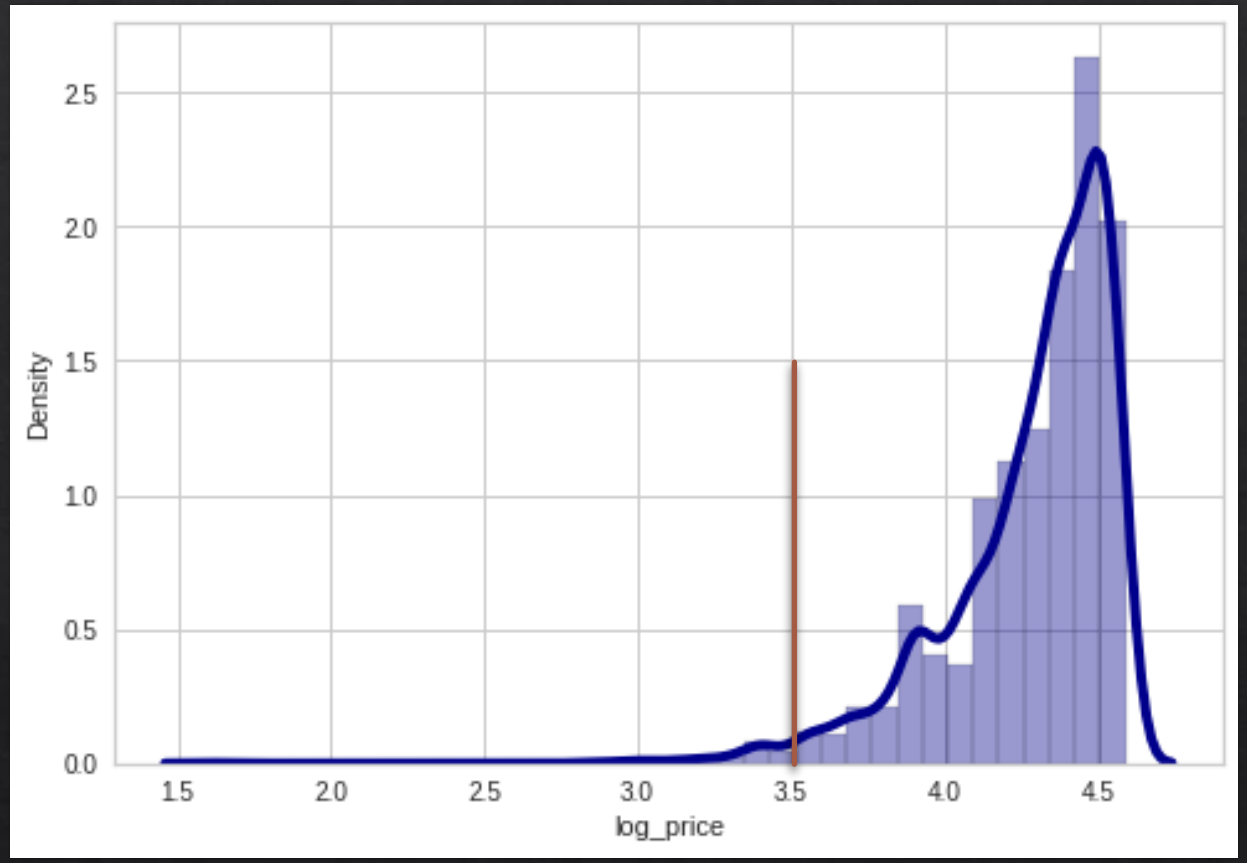


Outliers

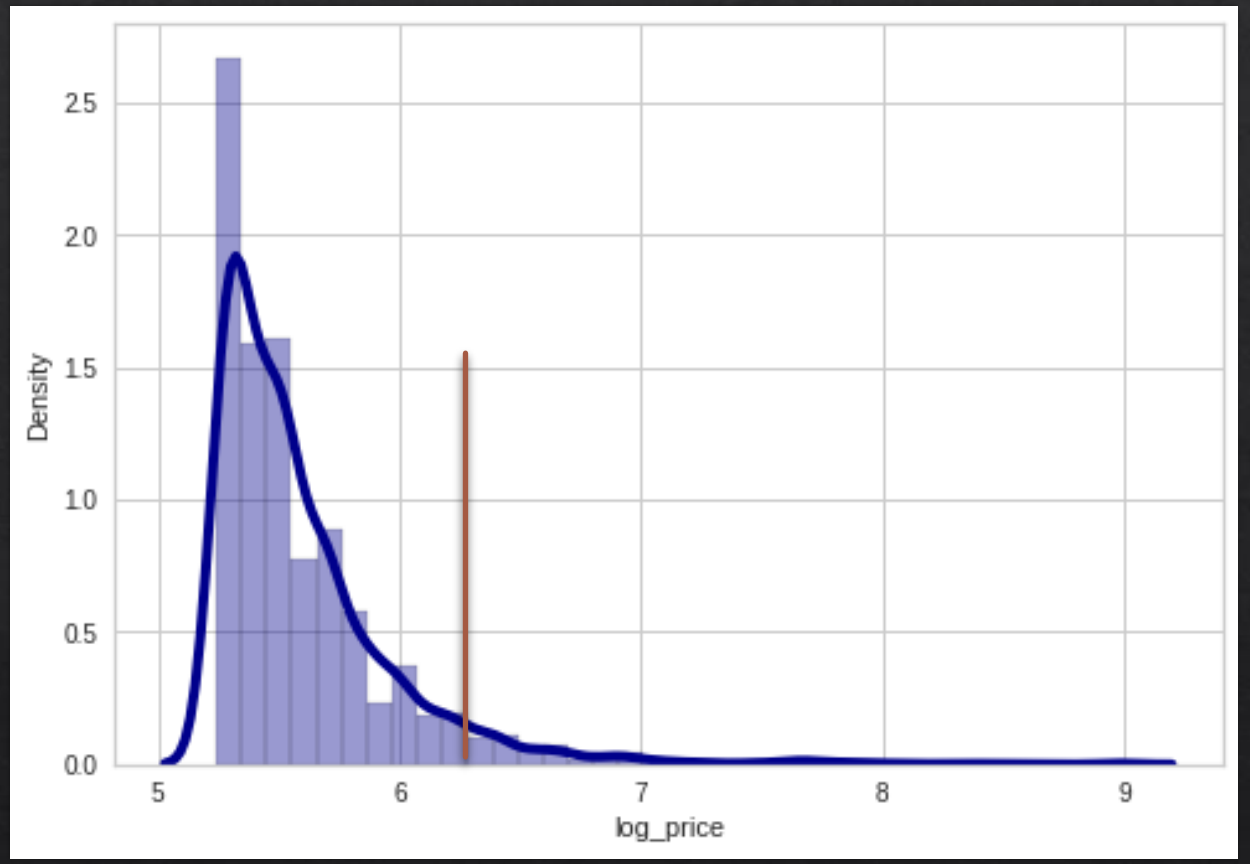
- 1st Quantile (4.59) : 4695 Values
- 2nd Quantile (4.90): 9389 Values
- 3rd Quantile (5.24): 14083 Values

- 4695 (25%) values under 4.59
- 9388 (50%) values between 4.59 – 5.24
- 4694 (25%) values over 5.24

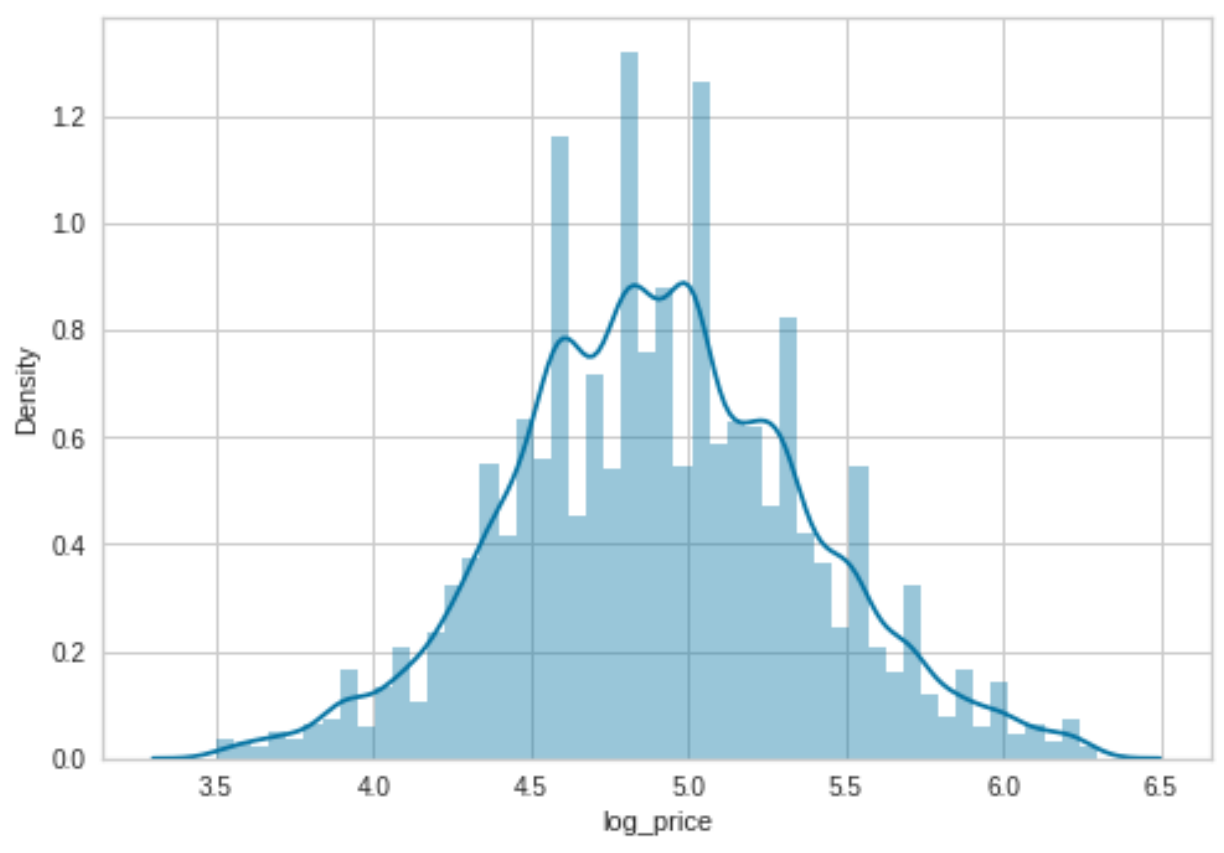
Cutting Outliers (<4.59)



Cutting Outliers (>5.24)

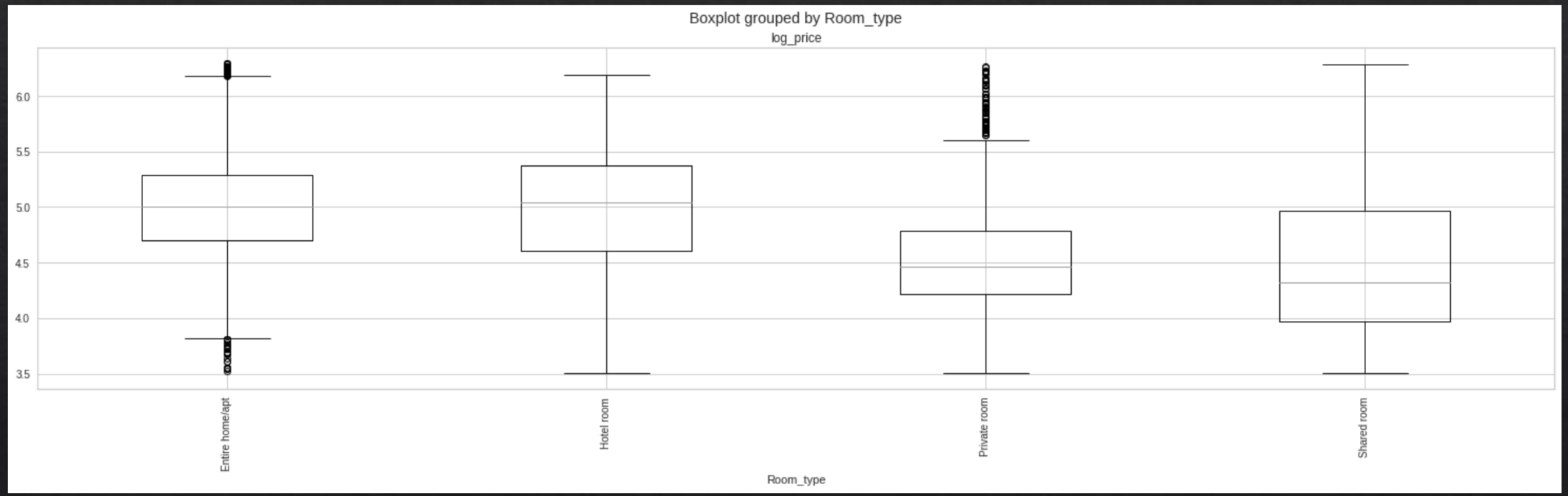


Log Price Distribution

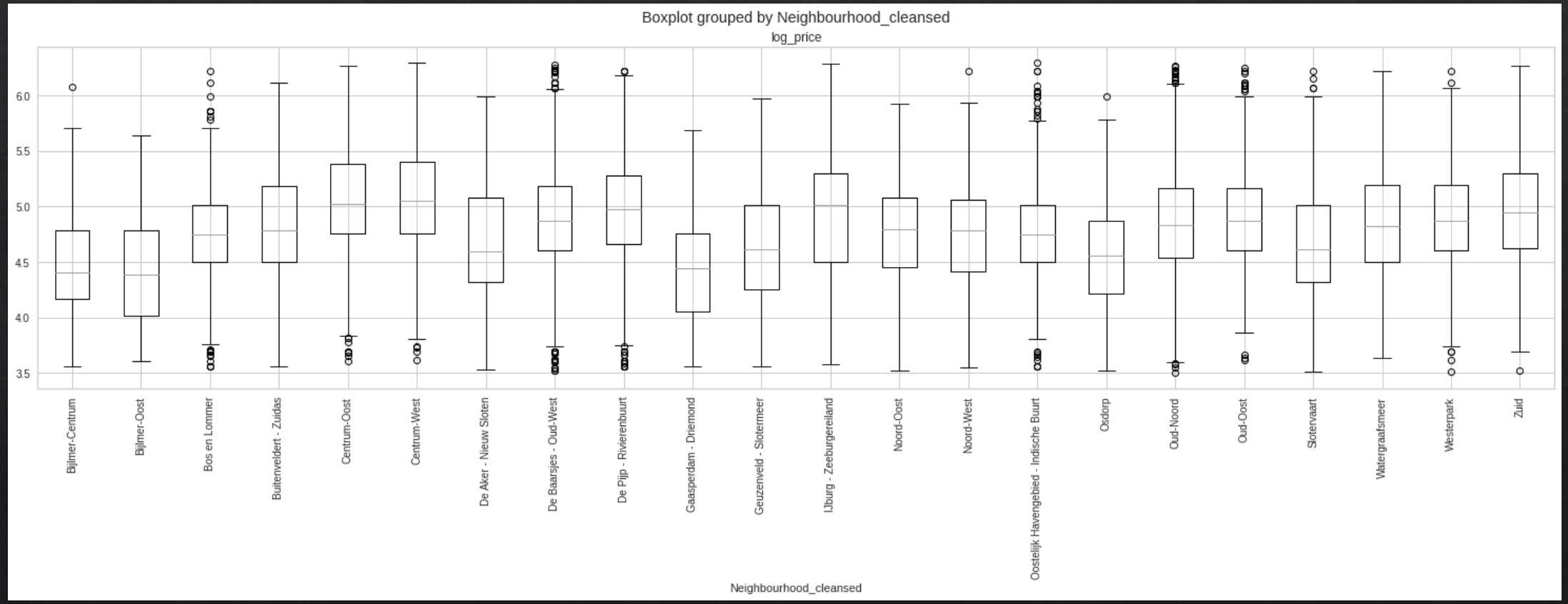


- Skewness = 0.092279
- Kurtosis = 0.046570
- Mean = 4.906654
- St. Deviation = 0.475776

Log Price vs Room Type

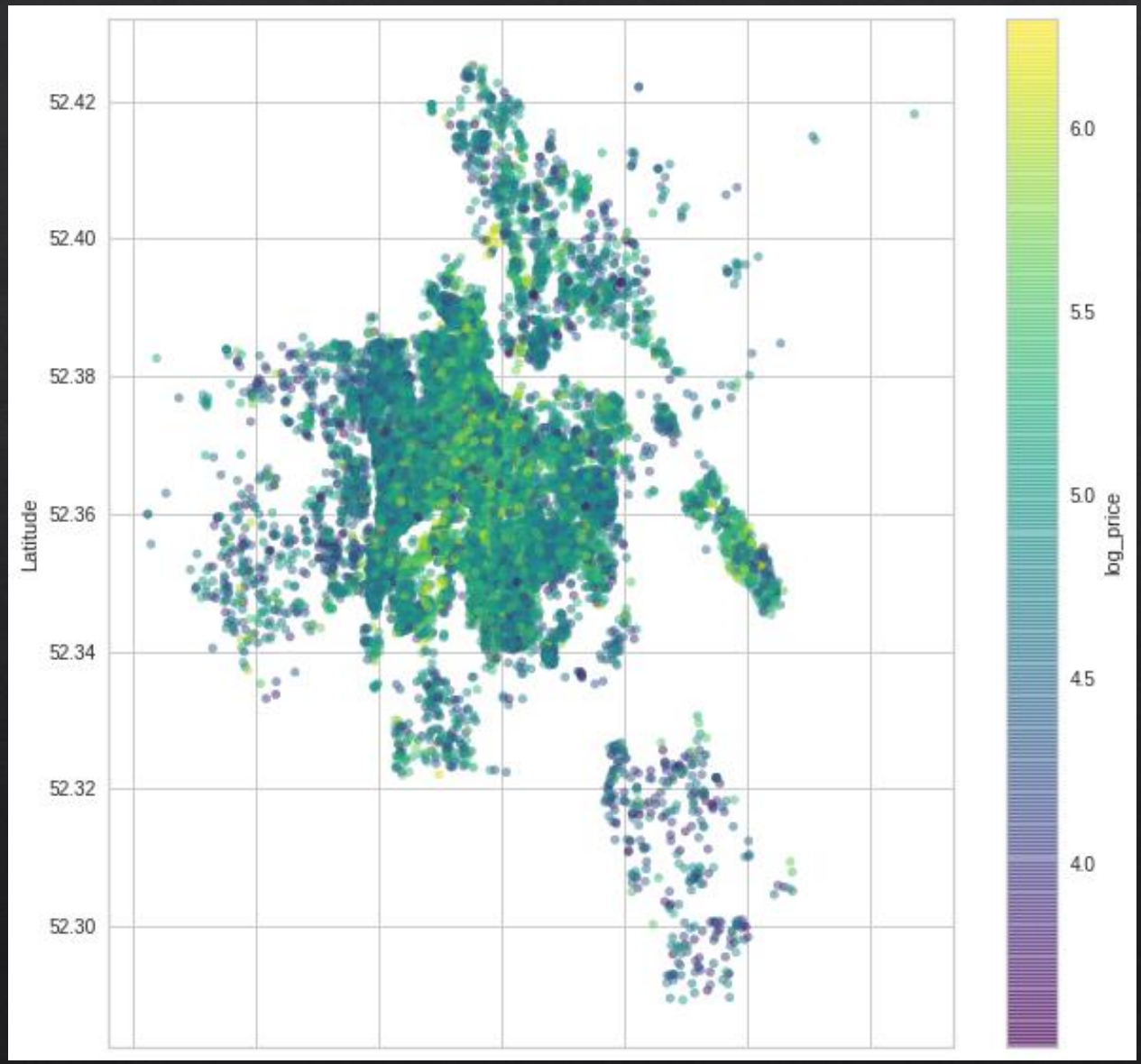


Log Price vs Neighborhood



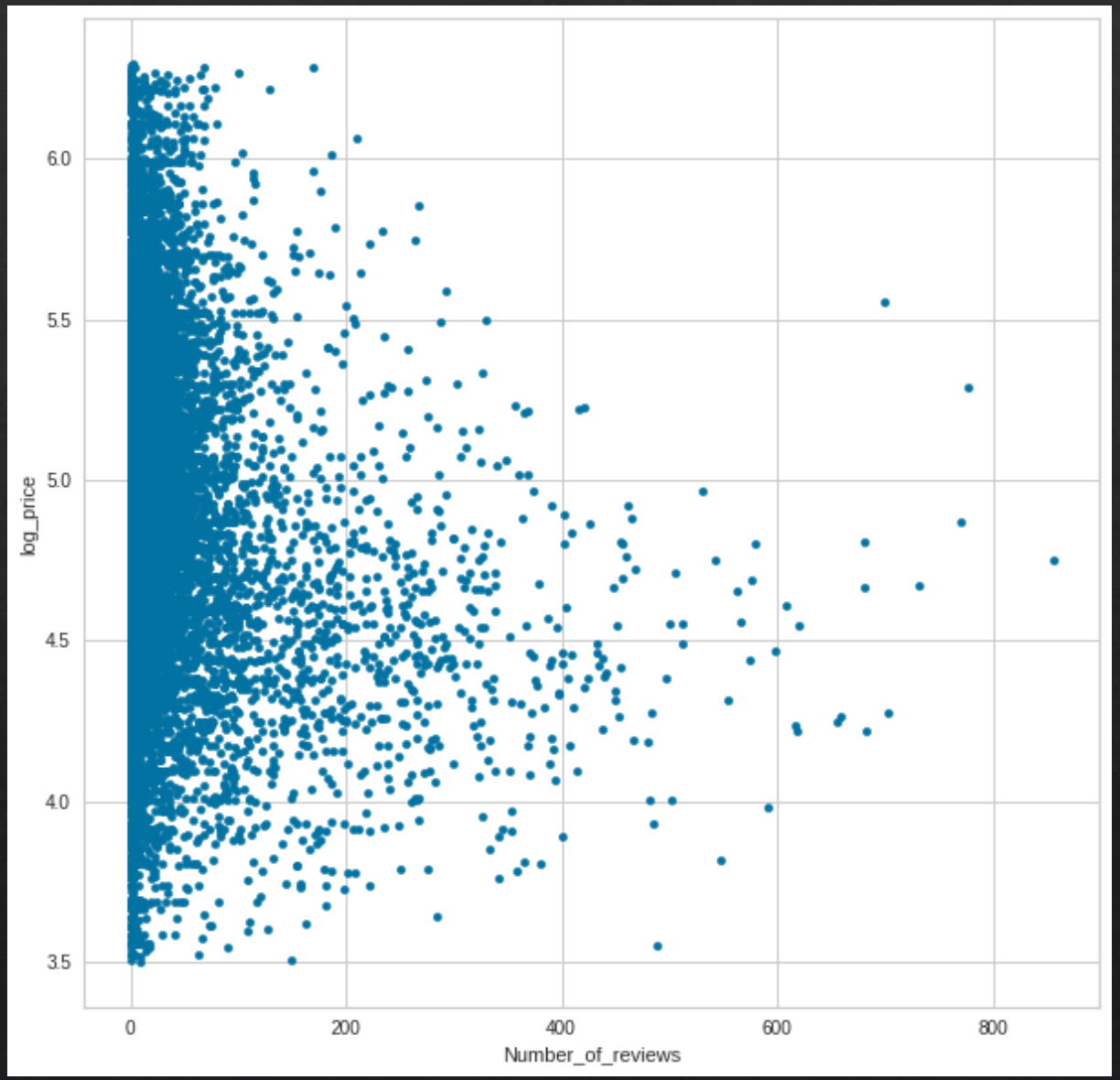
Log Price vs Location

- Expensive listings are mostly located in the center of the city
- Cheaper listings are in the suburbs



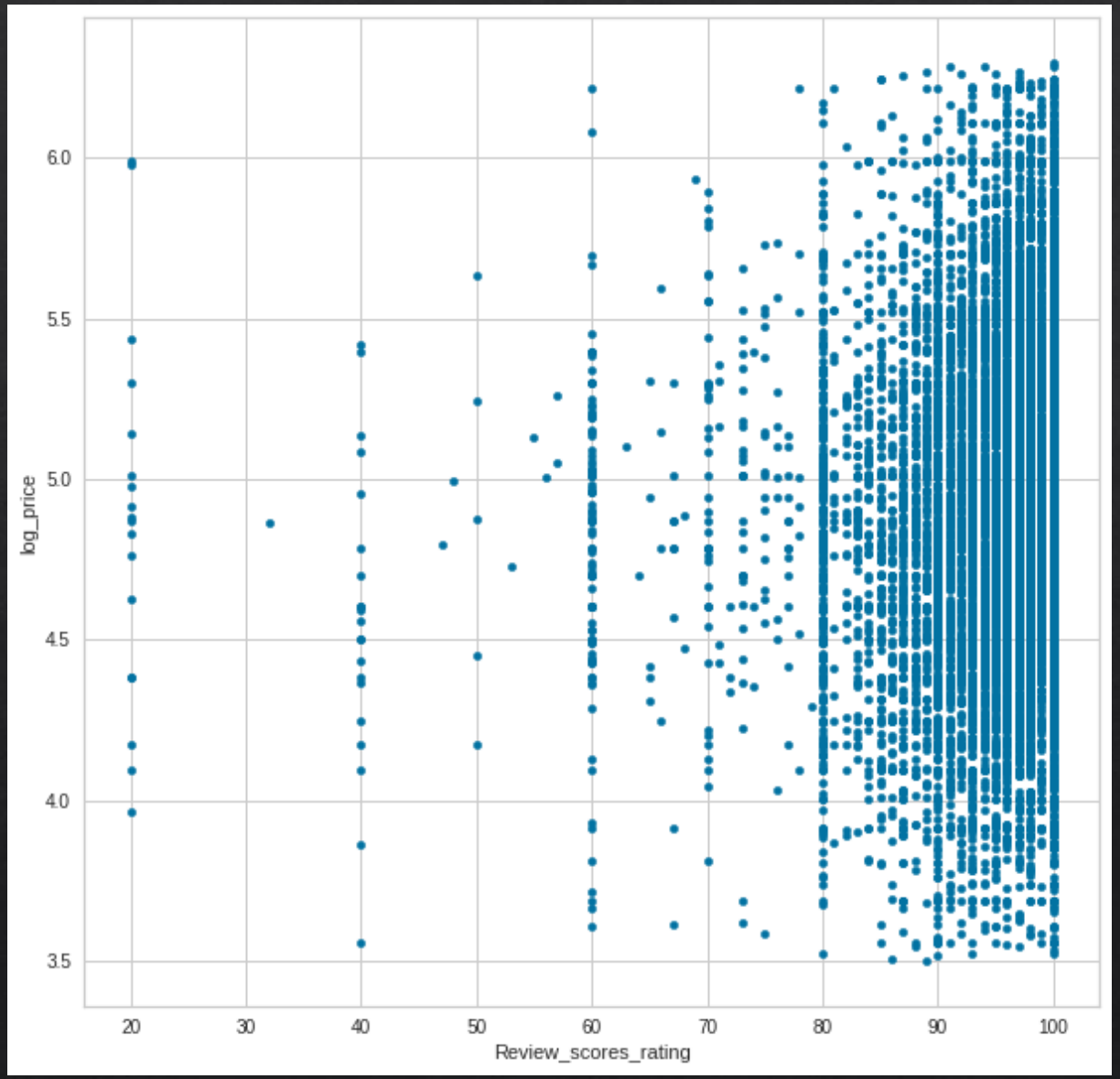
Log Price vs Number of Reviews

- Most listings have under 200 number of reviews
- More expensive listings do not have greater number of reviews



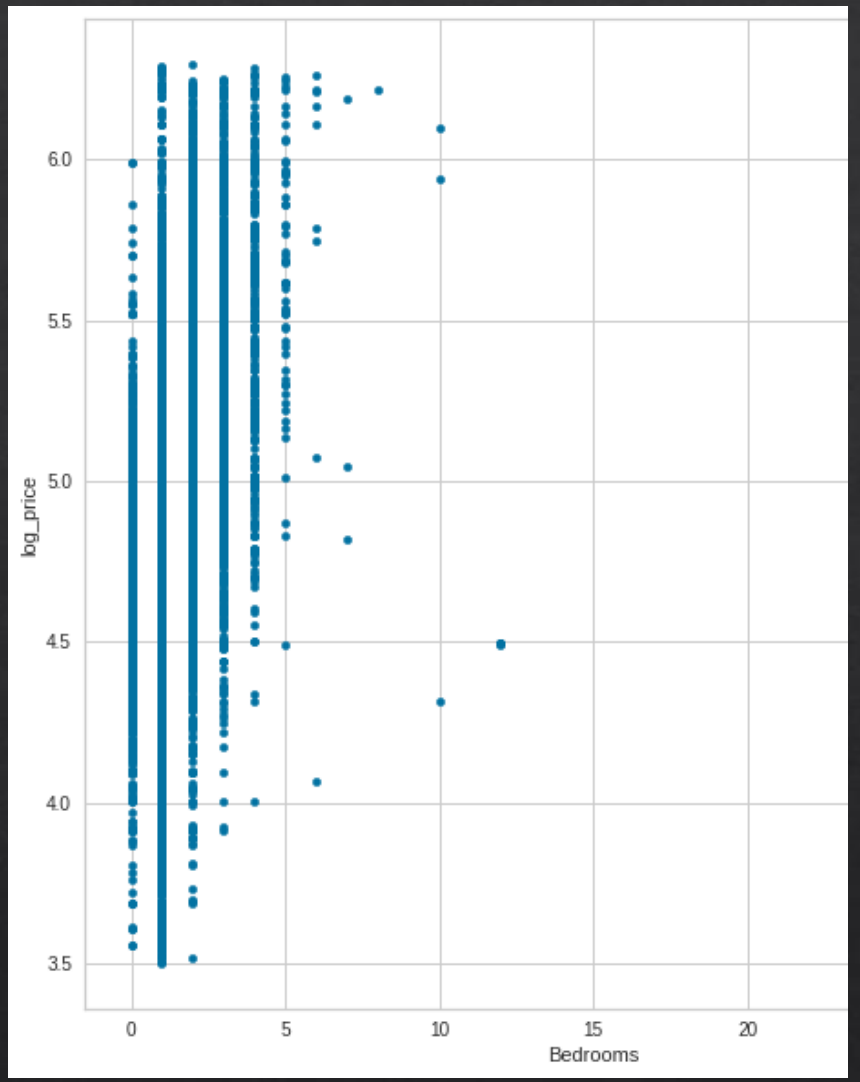
Log Price vs Score Rating

- Most listings have score rating over 80%
- Listings with bigger rating do not have bigger price



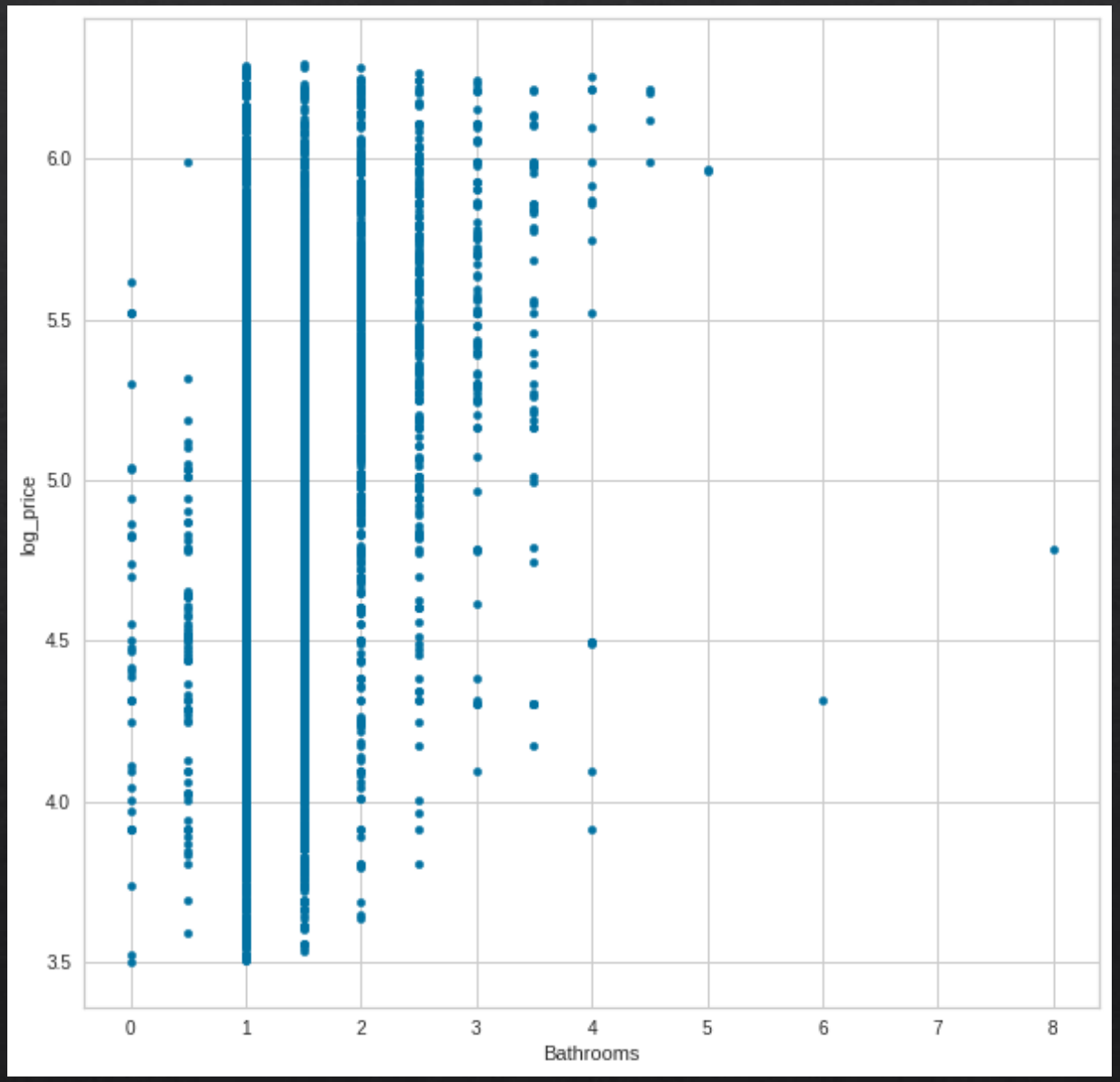
Log Price vs Bedrooms

- Most listings have under 5 bedrooms
- Listings with more bedrooms have higher price

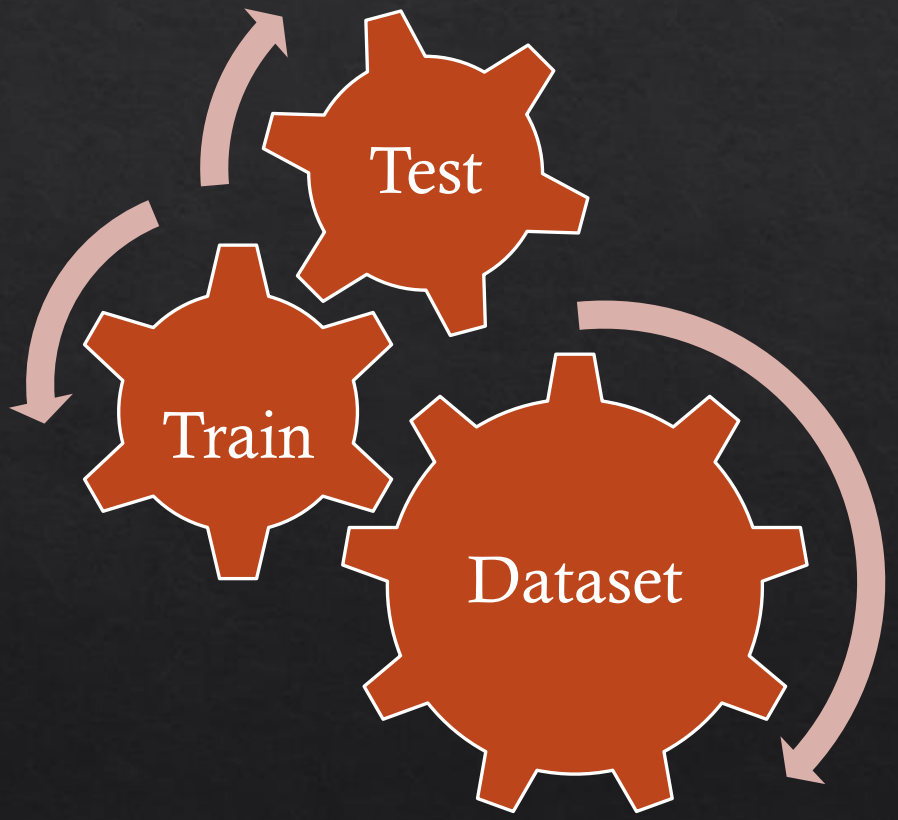


Log Price vs Bathrooms

- Most listings have under 4 bathrooms
- Listings with more bathrooms have higher price



Feature Engineering



Split into dependent and independent variable

Dataset				
Log_price	bathrooms	bedrooms	….	Amenity count
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .

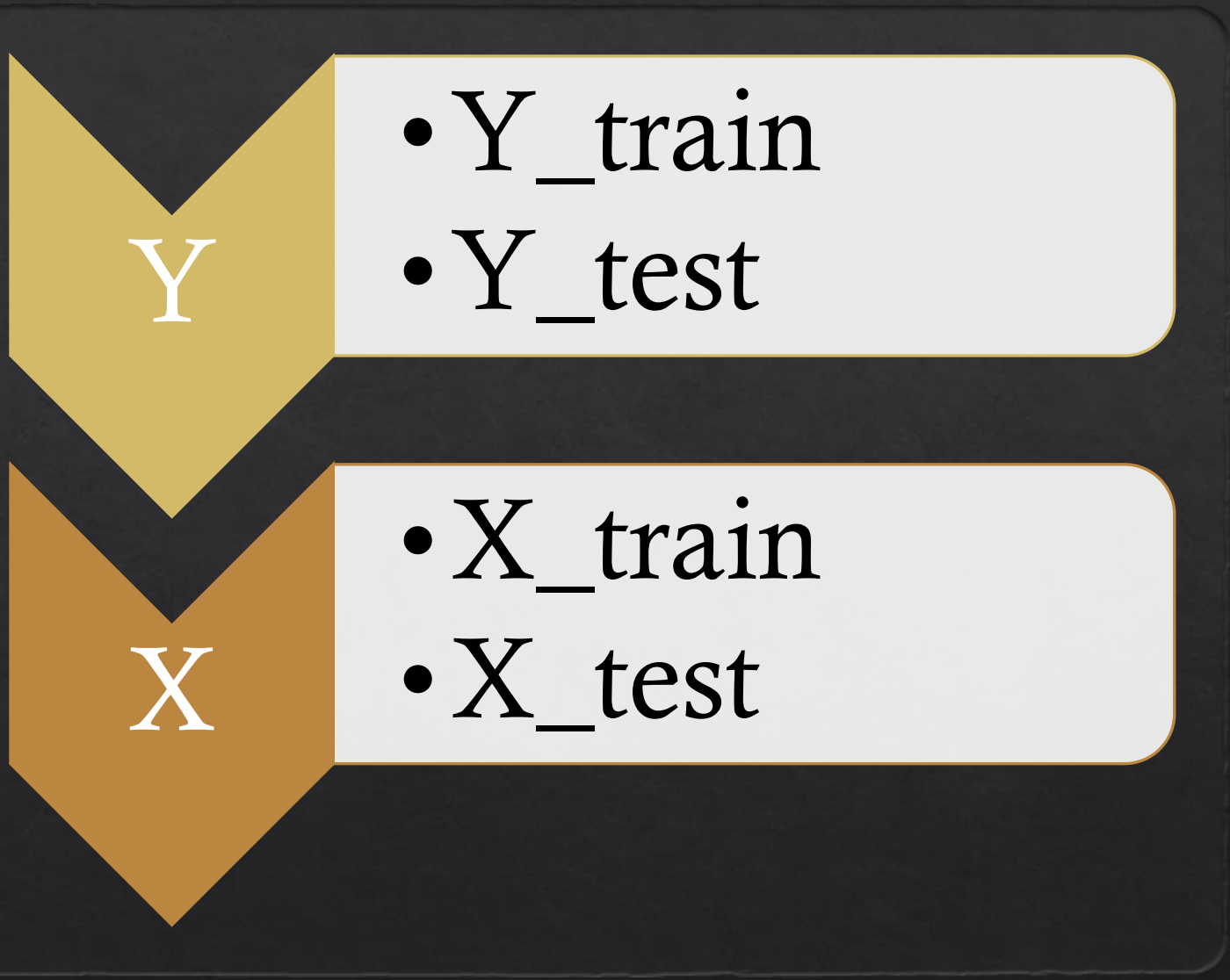
y
Log_price
. . .
. . .
. . .

x			
bathrooms	bedrooms	…	Amenity count
. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .

We choose to examine the price path based on the following regression models:

- ❑ Linear Regression
- ❑ Lasso
- ❑ Elastic Net
- ❑ Ridge
- ❑ Random Forest

*Baseline error: 0.478988

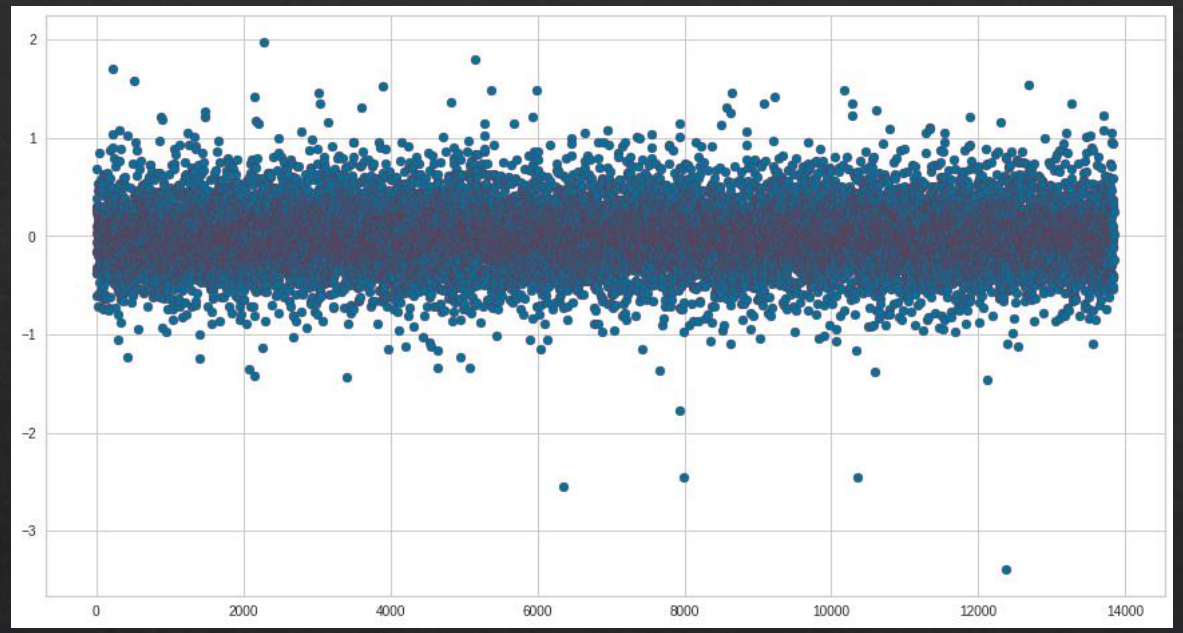


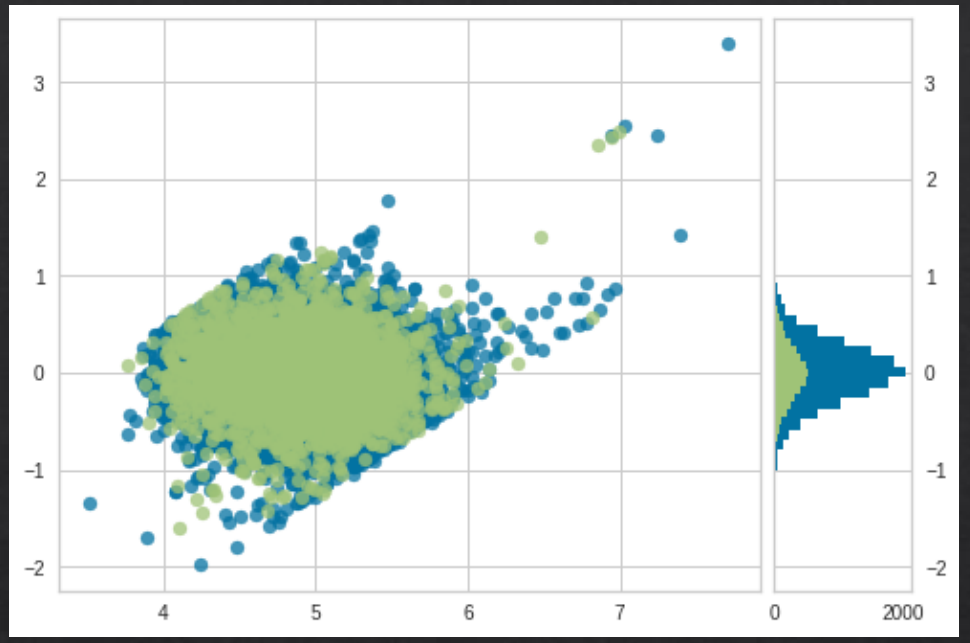
Model Prediction

20/12/2020

Linear Regression

Metric	Value
R^2	0.50415
Max. Error	2.49728
MSE	0.11345
MRSE	0.33683





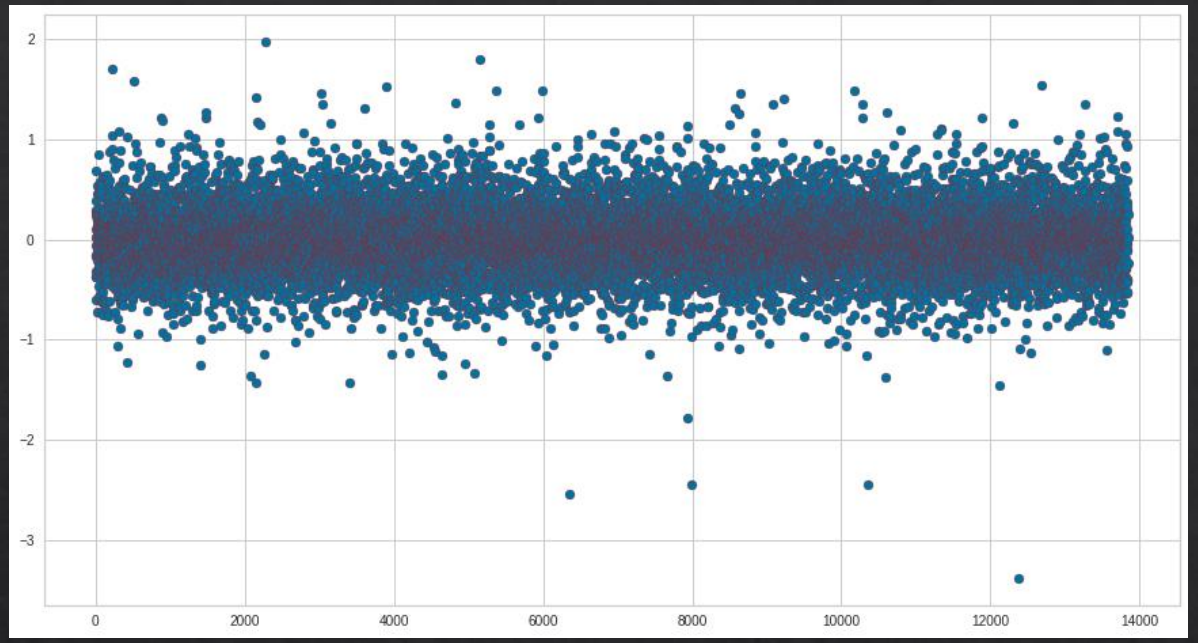
Residuals Scatter Plot

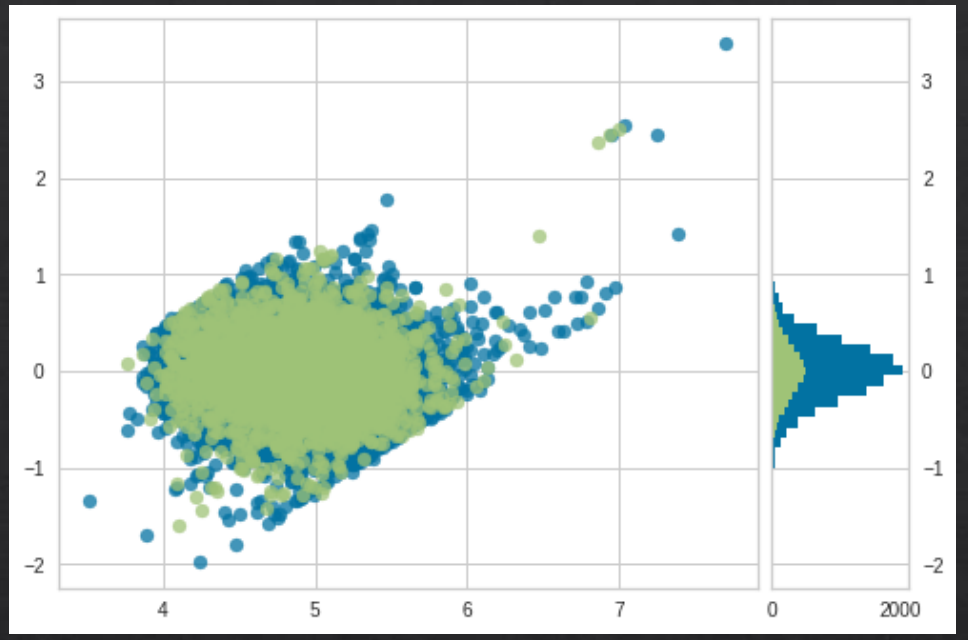
Residuals VS Predicted Price

20/12/2020

Ridge Regression

Metric	Value
R^2	0.50430
Max. Error	2.50040
MSE	0.11342
MRSE	0.33678



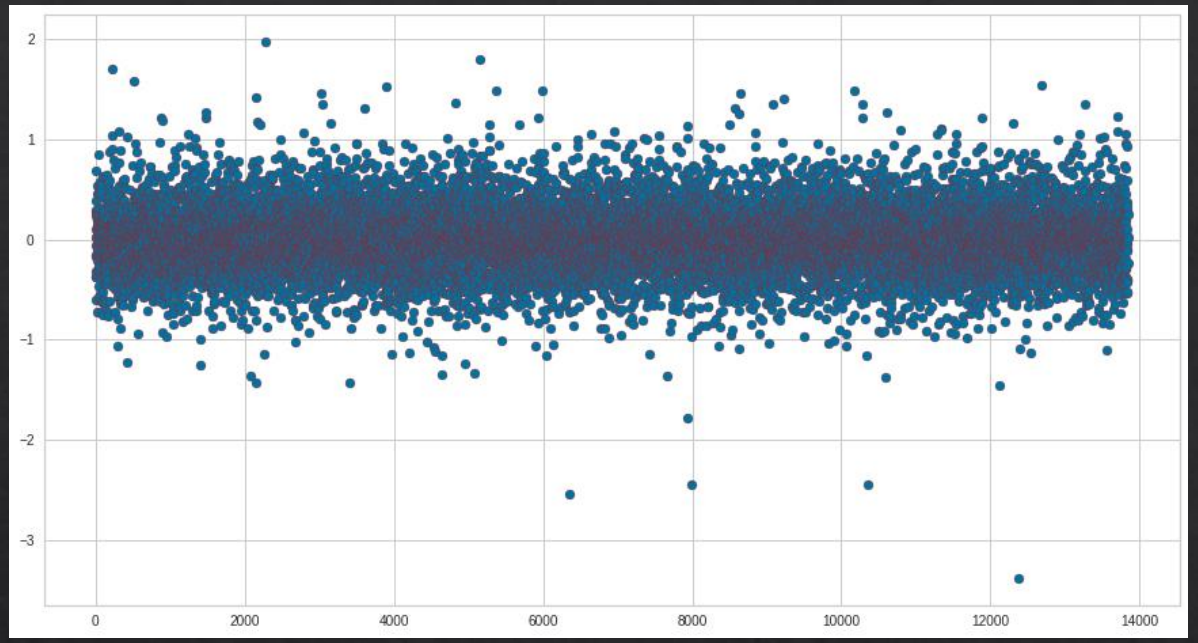


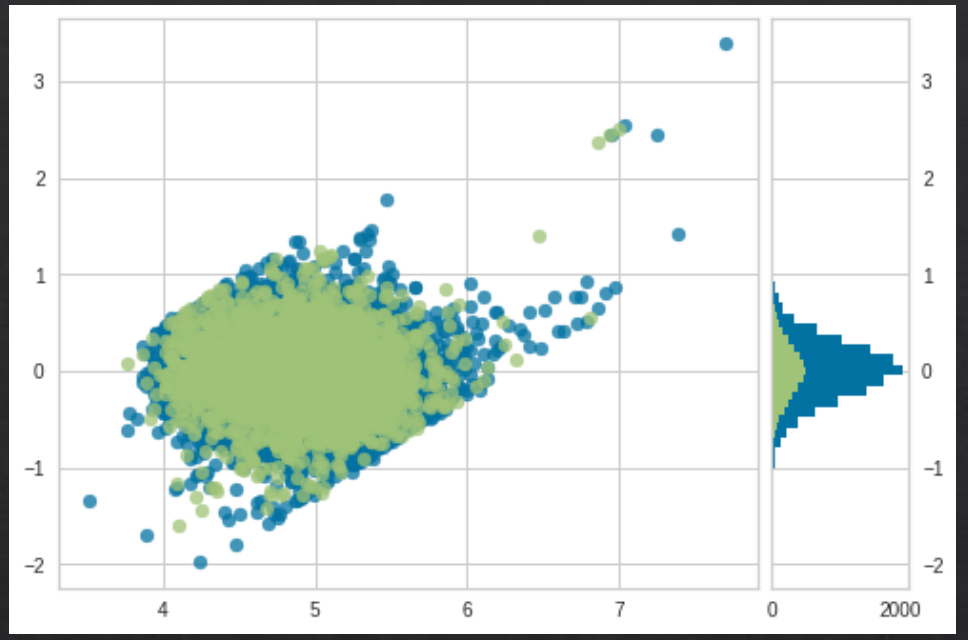
Residuals Scatter Plot

Residuals VS Predicted Price

Lasso Regression

Metric	Value
R^2	0.48356
Max. Error	2.51292
MSE	0.11816
MRSE	0.34374



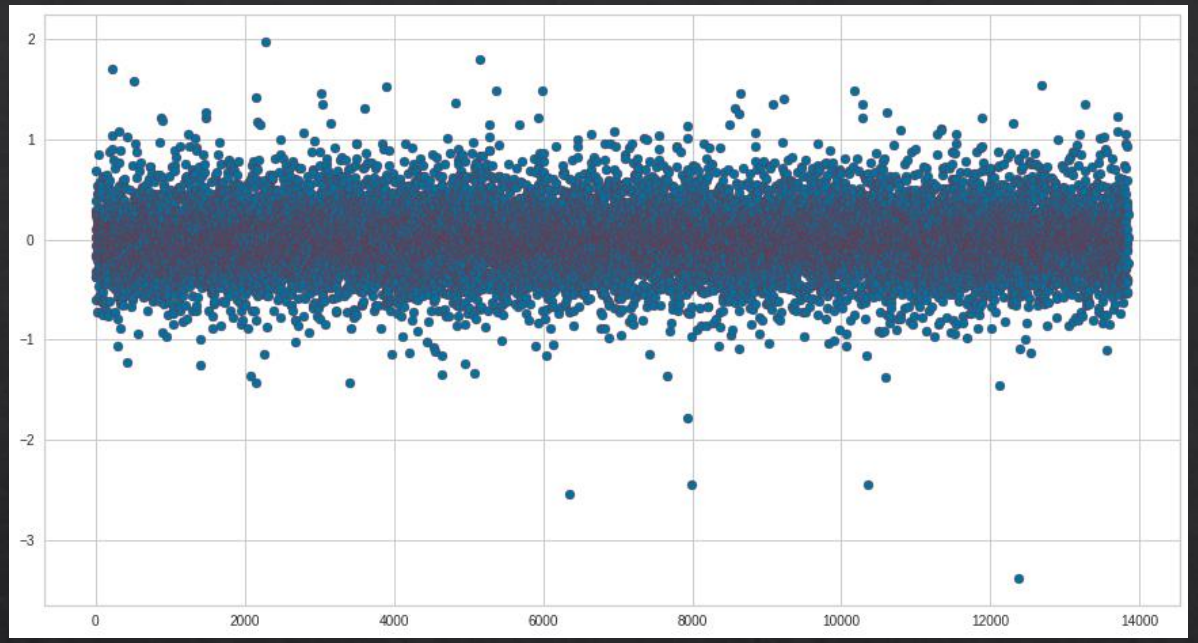


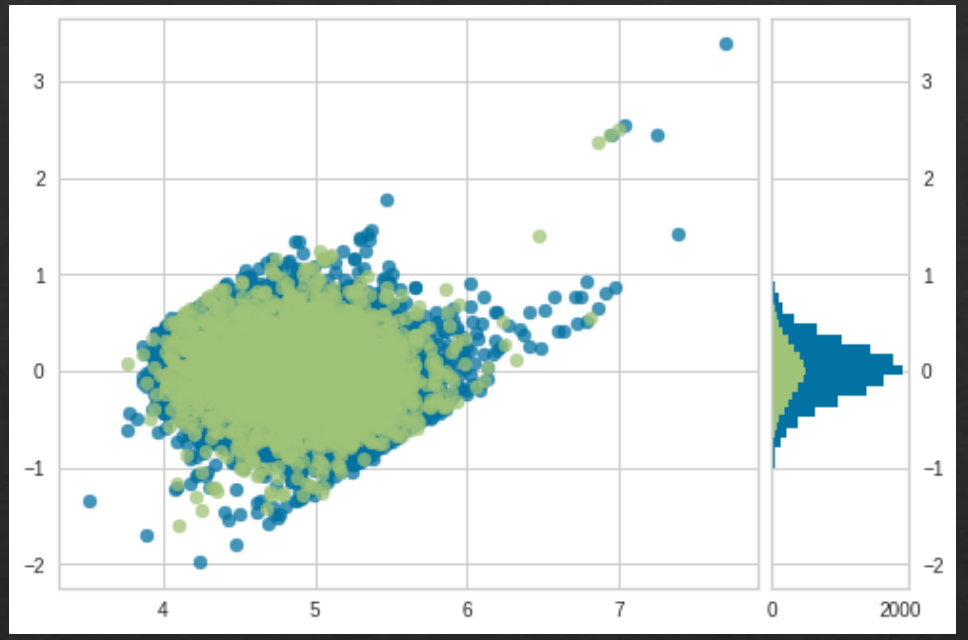
Residuals Scatter Plot

Residuals VS Predicted Price

Elastic Net Regression

Metric	Value
R^2	0.50415
Max. Error	2.52100
MSE	0.11345
MRSE	0.33683



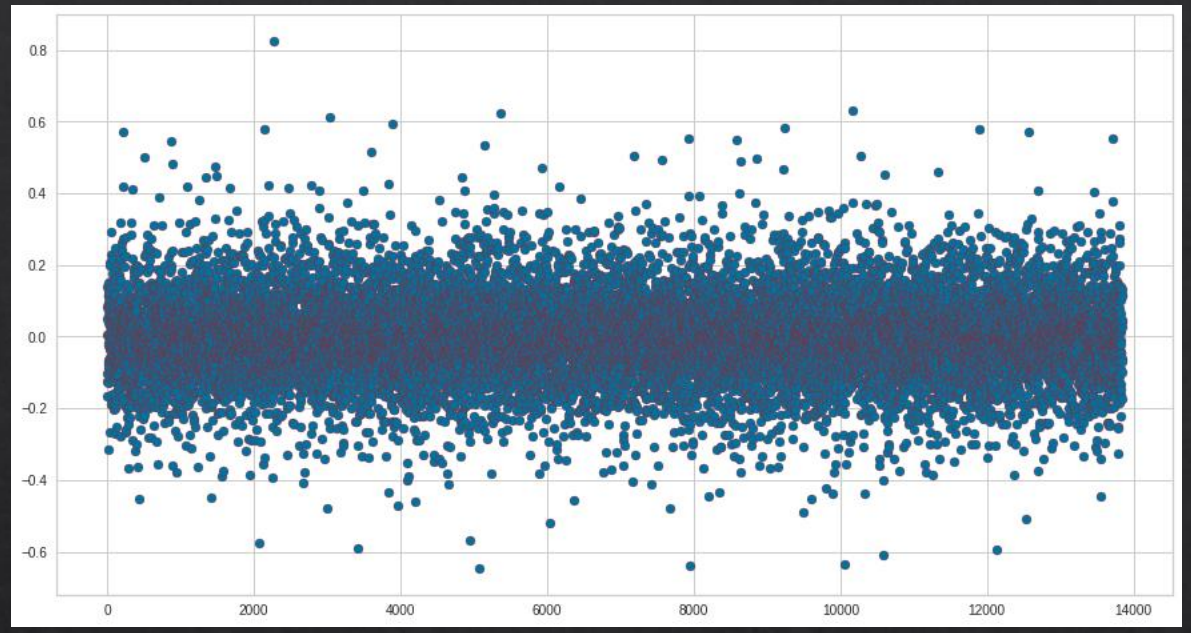


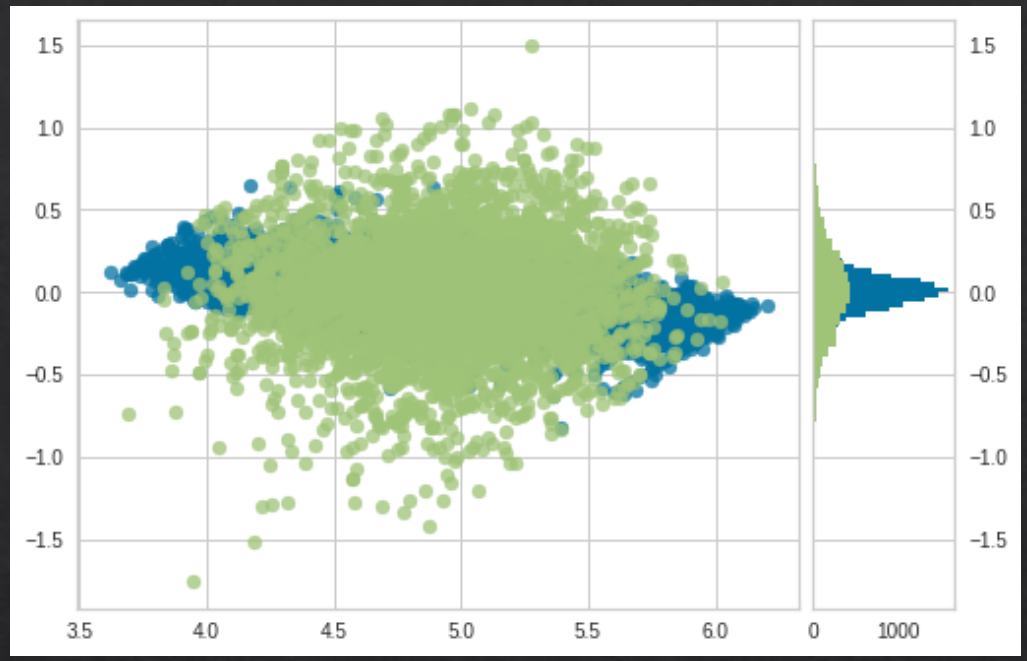
Residuals Scatter Plot

Residuals VS Predicted Price

Random Forest Regression

Metric	Value
R^2	0.54282
Max. Error	1.75878
MSE	0.10461
MRSE	0.32343



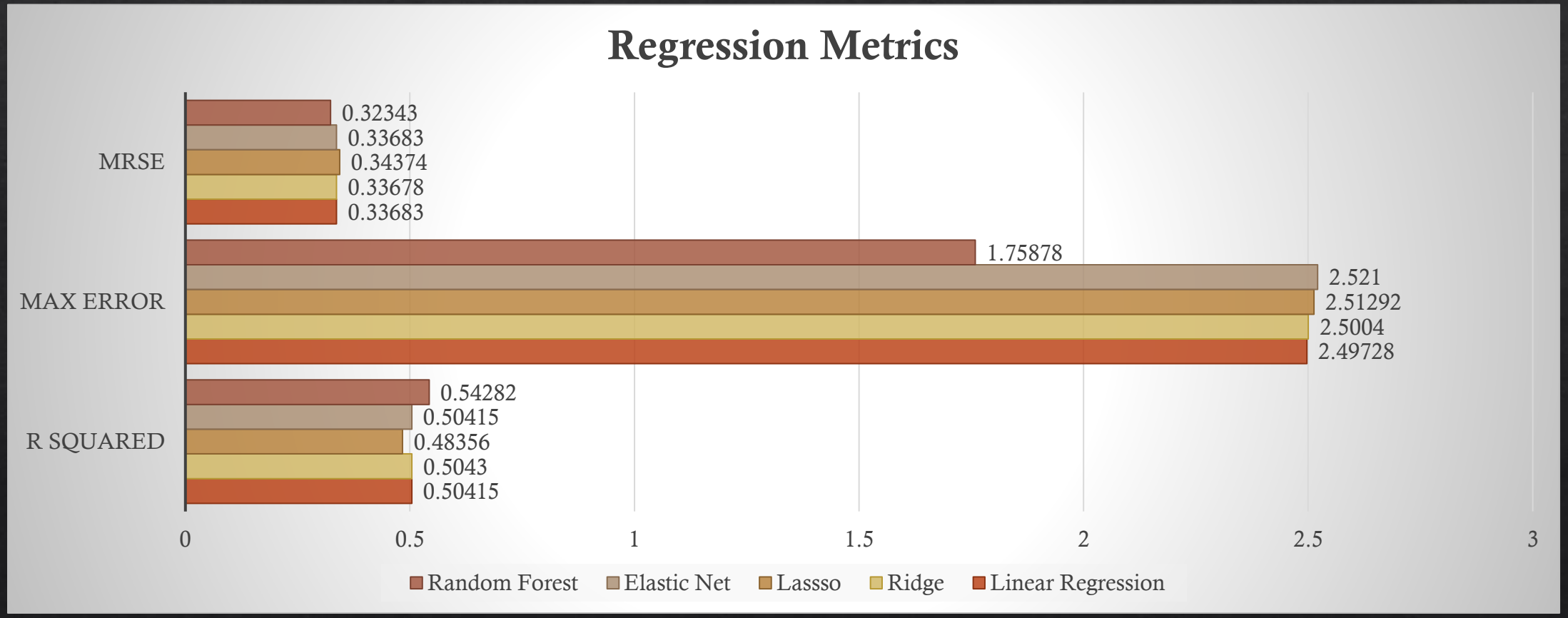


Residuals Scatter Plot

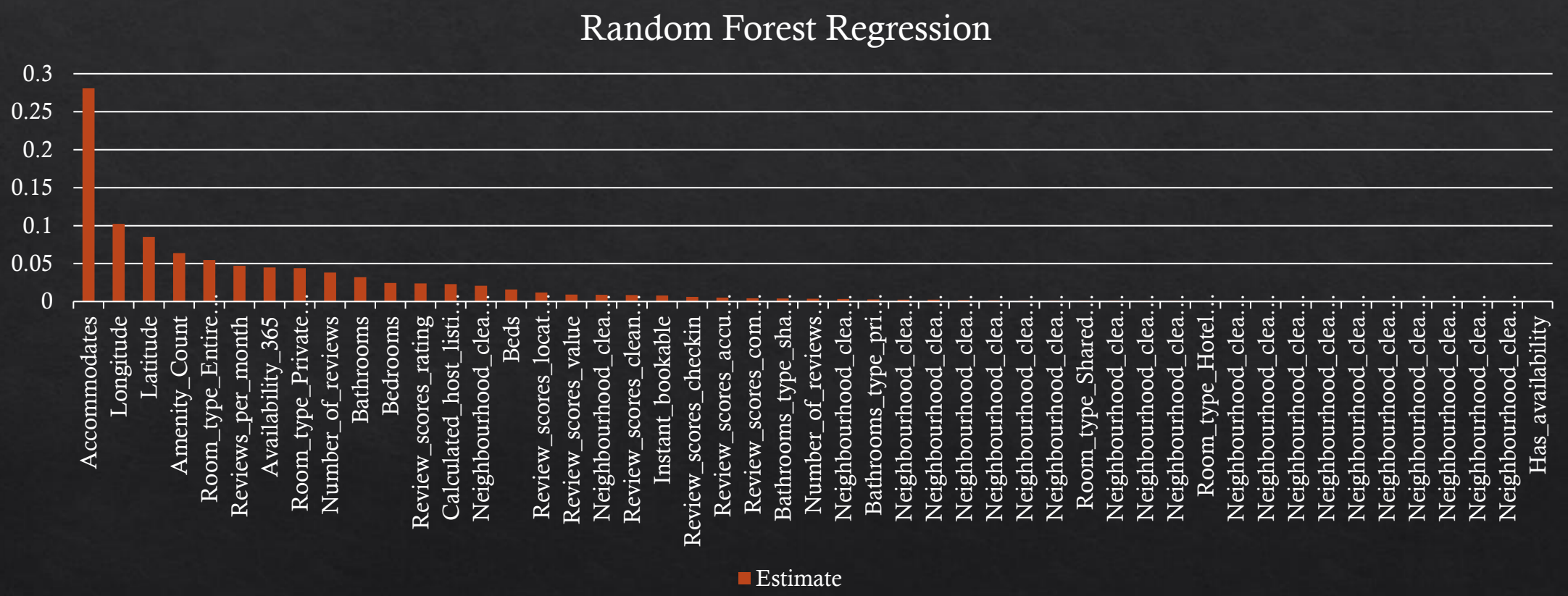
Residuals VS Predicted Price

20/12/2020

Model Evaluation

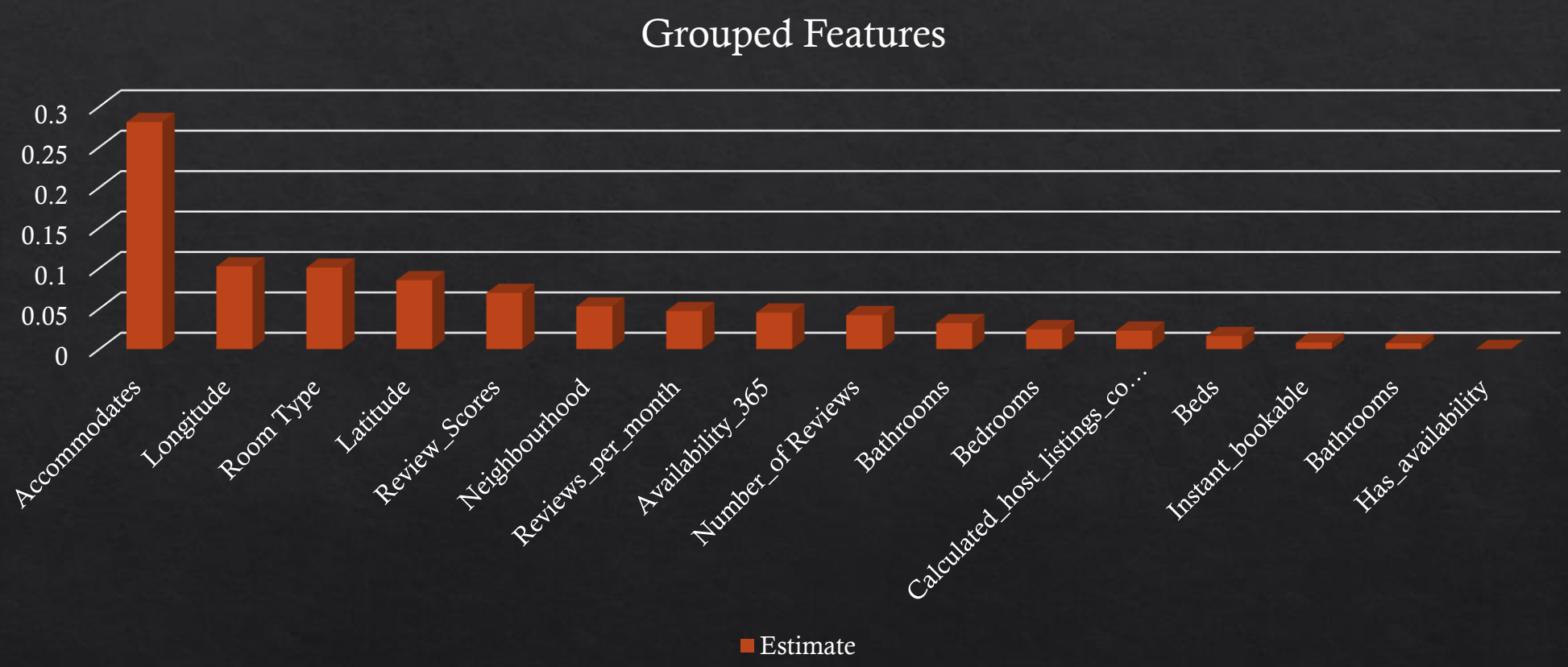


Feature Importance



20/12/2020

Feature Importance



20/12/2020