

# Descriptive Analysis and Correlation Analysis

## Overview


Dataset statistics	
Number of variables	6
Number of observations	38
Missing cells	39
Missing cells (%)	17.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.9 KiB
Average record size in memory	51.4 B

Variable types	
Categorical	2
Numeric	4

Alerts

Area is highly overall correlated with Average \$/sqft and 1 other fields (Average \$/sqft, listing_type)	High correlation
Average \$/sqft is highly overall correlated with Area	High correlation
listing_type is highly overall correlated with Area	High correlation
Available Units has 19 (50.0%) missing values	Missing
Price has 8 (21.1%) missing values	Missing
Area has 2 (5.3%) missing values	Missing
Average \$/sqft has 10 (26.3%) missing values	Missing

# Variables

Select Columns 

listing\_type  
Categorical

Distinct	3
Distinct (%)	7.9%
Missing	0
Missing (%)	0.0%
Memory size	432.0 B

## Length

Max length	1
Median length	1
Mean length	1
Min length	1

## Characters and Unicode

Total characters	38	
Distinct characters	3	
Distinct categories	1 ( <a href="https://en.wikipedia.org/wiki/Unicode_character_property#General_Category">https://en.wikipedia.org/wiki/Unicode_character_property#General_Category</a> )	?
Distinct scripts	1 ( <a href="https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode">https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode</a> )	?
Distinct blocks	1 ( <a href="https://en.wikipedia.org/wiki/Unicode_block">https://en.wikipedia.org/wiki/Unicode_block</a> )	?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

## Unique

Unique	0	?
Unique (%)	0.0%	

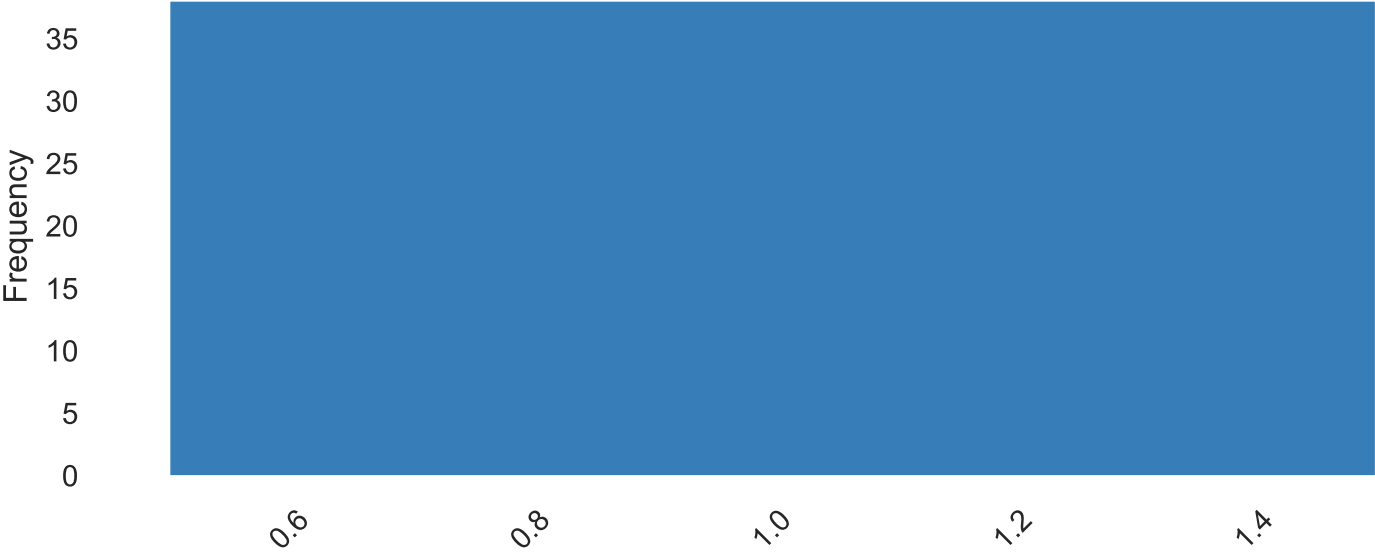
## Sample

1st row	1
2nd row	1
3rd row	1
4th row	1
5th row	1

Common Values

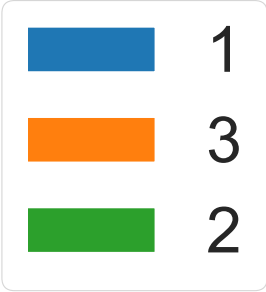
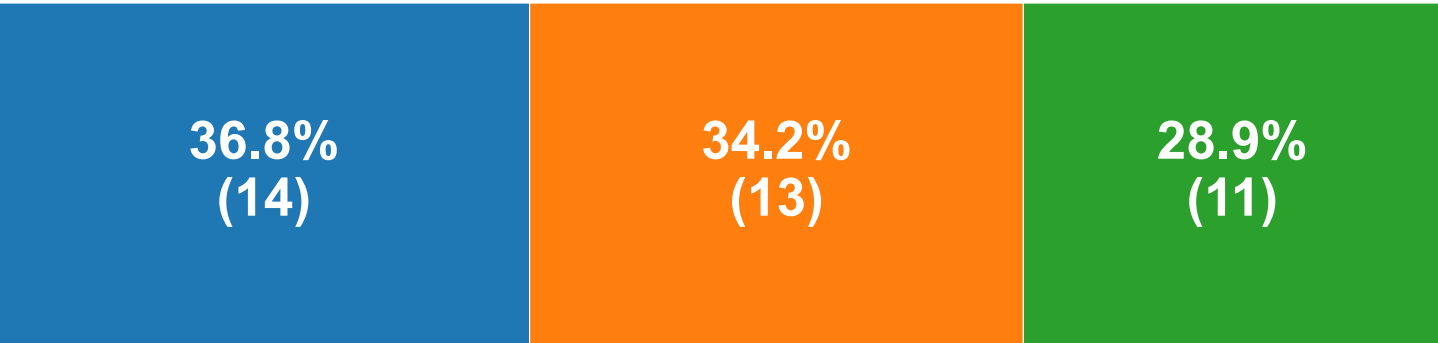
Value	Count	Frequency (%)
1	14	36.8%
3	13	34.2%
2	11	28.9%

Length



Histogram of lengths of the category

Common Values (Plot)



Name

Categorical

Distinct	19
Distinct (%)	50.0%
Missing	0
Missing (%)	0.0%
Memory size	432.0 B

Length

Max length	30
Median length	21
Mean length	18.078947
Min length	11

Characters and Unicode

Total characters	687	
Distinct characters	49	
Distinct categories	5 ( <a href="https://en.wikipedia.org/wiki/Unicode_character_property#General_Category">https://en.wikipedia.org/wiki/Unicode_character_property#General_Category</a> )	?
Distinct scripts	2 ( <a href="https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode">https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode</a> )	?
Distinct blocks	1 ( <a href="https://en.wikipedia.org/wiki/Unicode_block">https://en.wikipedia.org/wiki/Unicode_block</a> )	?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	8	?
Unique (%)	21.1%	

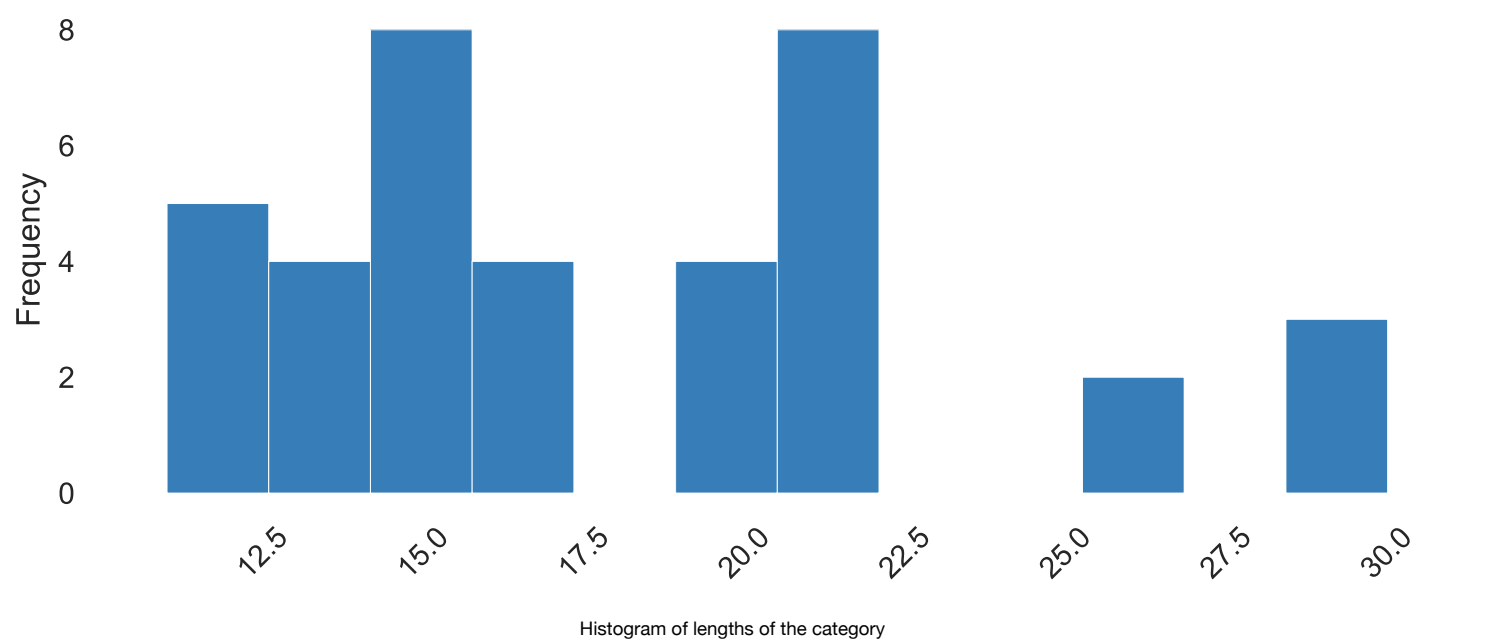
Sample

1st row	Madison Gardens
2nd row	Main Street Apartments
3rd row	Laurel at Dry Creek
4th row	Mosby at Bridgestreet
5th row	Willow Run Apartments

Common Values

Value	Count	Frequency (%)
Madison Gardens	3	7.9%
Emerald Ridge	3	7.9%
Laurel at Dry Creek	3	7.9%
Mosby at Bridgestreet	3	7.9%
The Paddock Club at Providence	3	7.9%
Highland Pointe	3	7.9%
Addison Park	3	7.9%
Main Street Apartments	3	7.9%
Royal Pines	2	5.3%
201 Pumprock Dr	2	5.3%
Other values (9)	10	26.3%

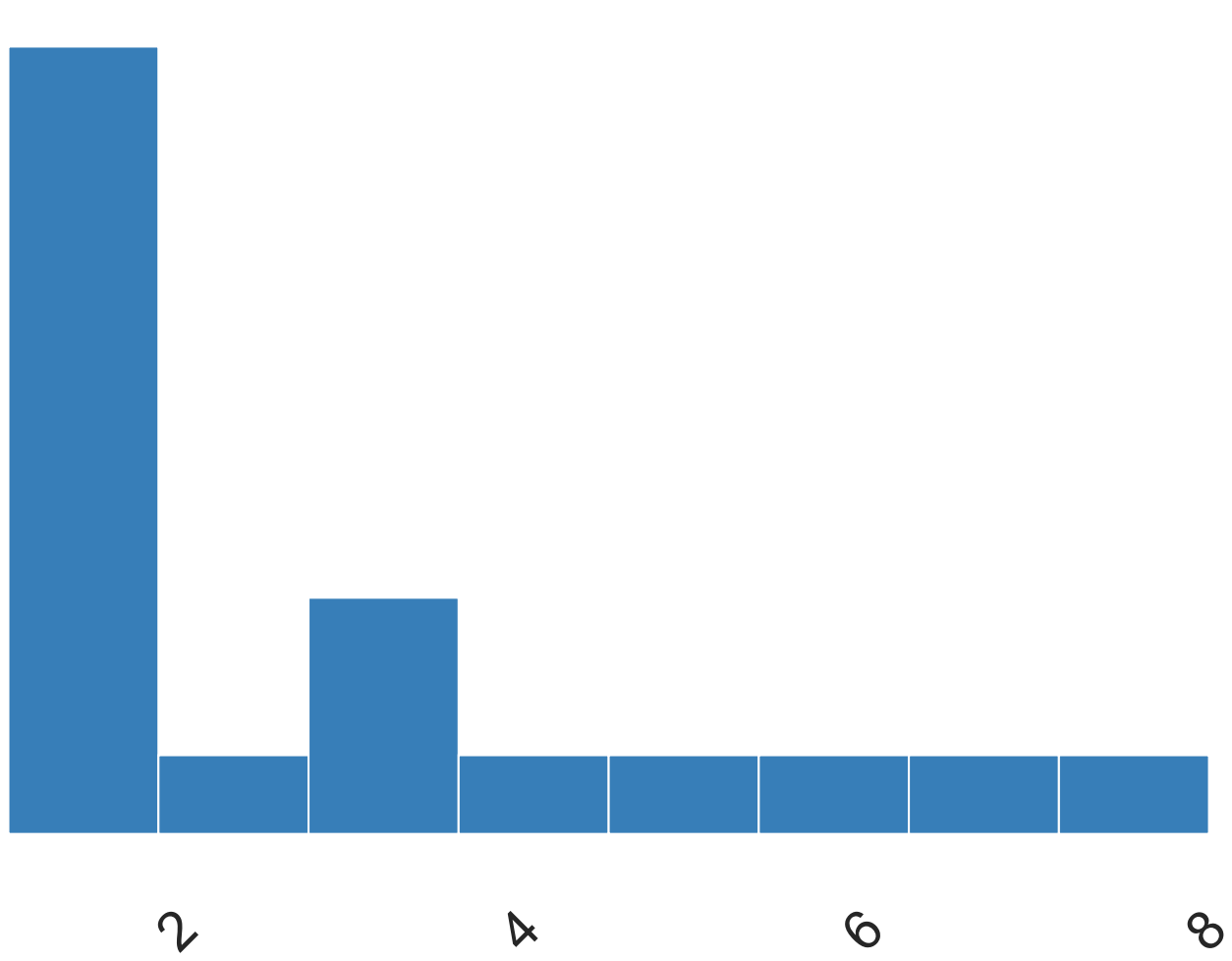
Length



Available Units

Real number ( $\mathbb{R}$ )

Distinct	8
Distinct (%)	42.1%
Missing	19
Missing (%)	50.0%
Infinite	0
Infinite (%)	0.0%
Mean	2.6842105
Minimum	1
Maximum	8
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	432.0 B



Quantile statistics

Minimum	1
5-th percentile	1
Q1	1
median	1
Q3	3.5

95-th percentile	7.1
Maximum	8
Range	7
Interquartile range (IQR)	2.5

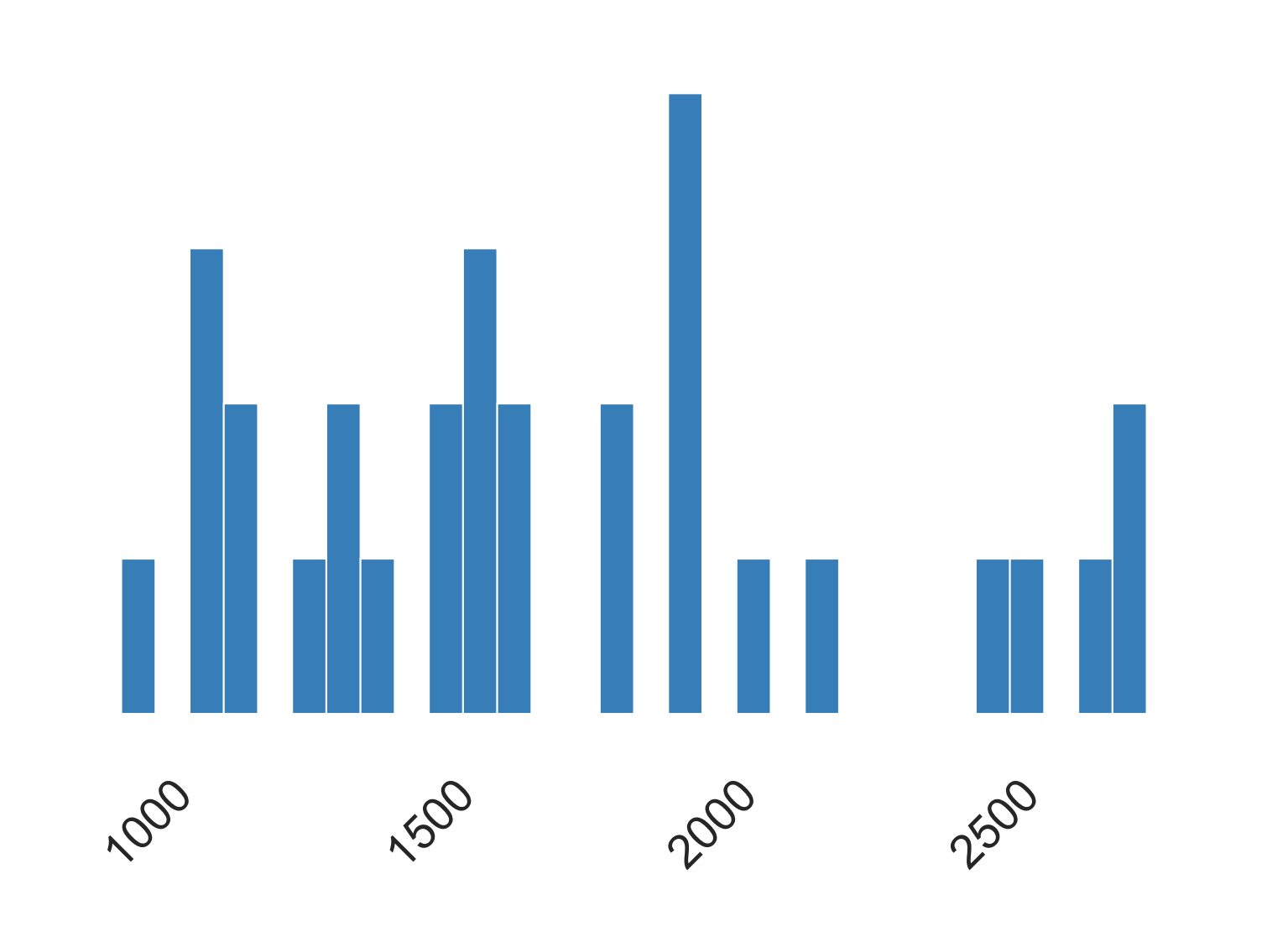
Descriptive statistics

Standard deviation	2.2864974
Coefficient of variation (CV)	0.85183235
Kurtosis	0.34164101
Mean	2.6842105
Median Absolute Deviation (MAD)	0
Skewness	1.2141517
Sum	51
Variance	5.2280702
Monotonicity	Not monotonic



Price  
Real number (ℝ)

Distinct	30
Distinct (%)	100.0%
Missing	8
Missing (%)	21.1%
Infinite	0
Infinite (%)	0.0%
Mean	1731.4
Minimum	950
Maximum	2765.5
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	432.0 B



Quantile statistics

Minimum	950
5-th percentile	1106.3
Q1	1336.625
median	1625
Q3	1966.125

95-th percentile	2713.75
Maximum	2765.5
Range	1815.5
Interquartile range (IQR)	629.5

Descriptive statistics

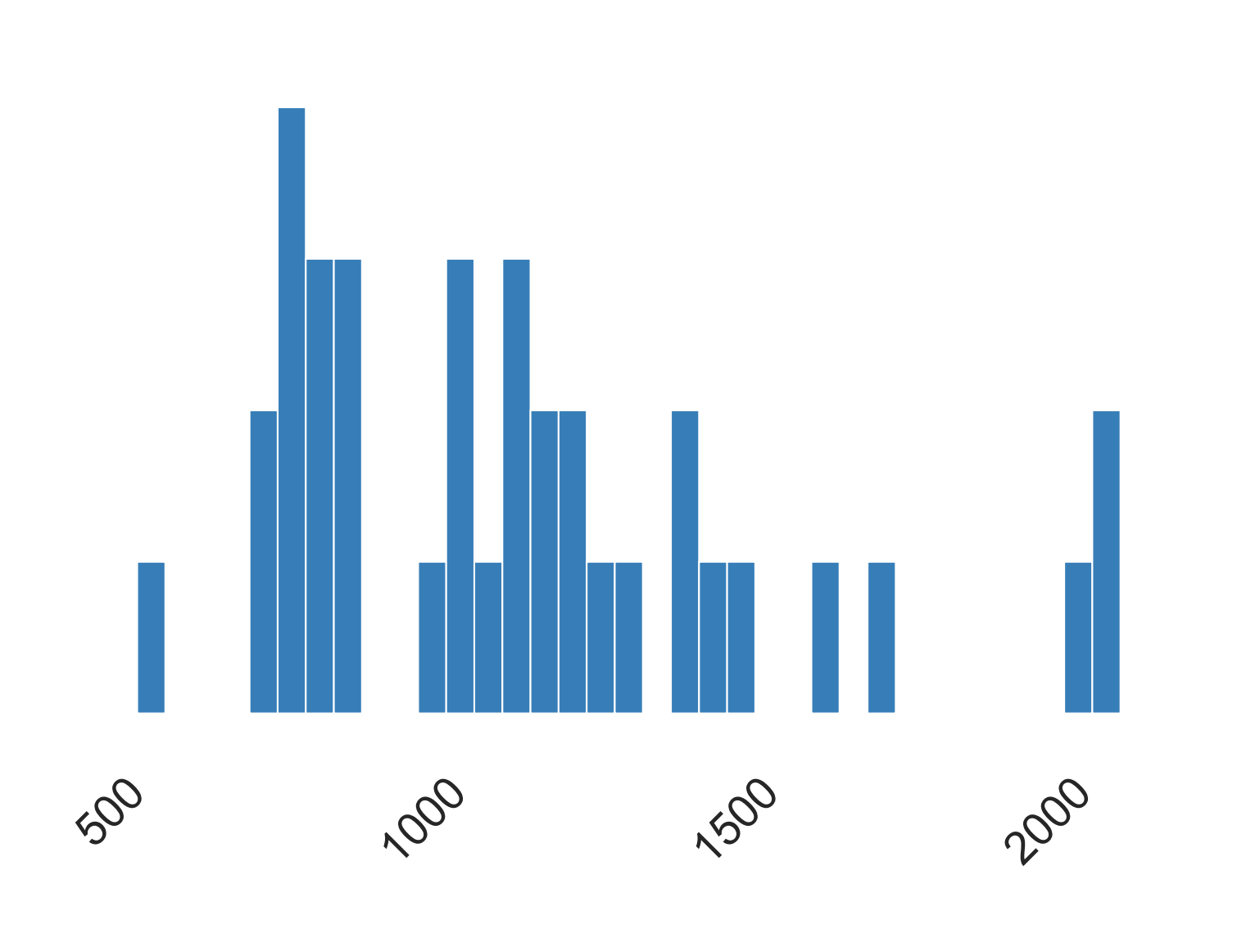
Standard deviation	523.29649
Coefficient of variation (CV)	0.30223893
Kurtosis	-0.55243794
Mean	1731.4
Median Absolute Deviation (MAD)	335.75
Skewness	0.57850972
Sum	51942
Variance	273839.21
Monotonicity	Not monotonic

Area

Real number (ℝ)

HIGH CORRELATION (This variable has a high overall correlation with 2 fields: Average \$/soft, listing type) MISSING

Distinct	35
Distinct (%)	97.2%
Missing	2
Missing (%)	5.3%
Infinite	0
Infinite (%)	0.0%
Mean	1166.7917
Minimum	540
Maximum	2113
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	432.0 B



Quantile statistics

Minimum	540
5-th percentile	747.75
Q1	850
median	1113.75

Q3	1356.875
95-th percentile	2045.75
Maximum	2113
Range	1573
Interquartile range (IQR)	506.875

Descriptive statistics

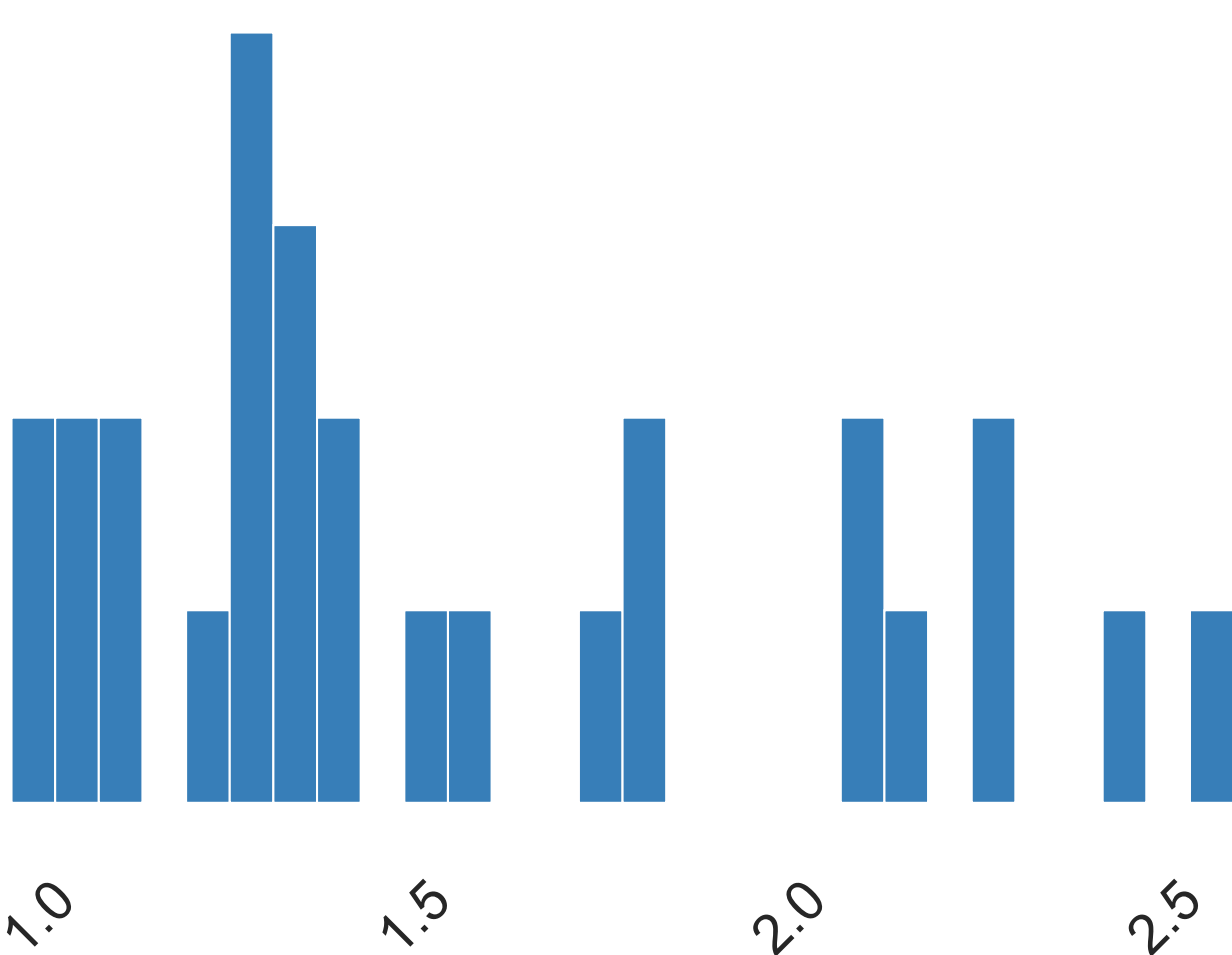
Standard deviation	393.19518
Coefficient of variation (CV)	0.33698833
Kurtosis	0.48268532
Mean	1166.7917
Median Absolute Deviation (MAD)	263.75
Skewness	0.95360146
Sum	42004.5
Variance	154602.45
Monotonicity	Not monotonic

Average \$/sqft

Real number (ℝ)

HIGH CORRELATION (This variable has a high overall correlation with 1 fields: Area) MISSING

Distinct	28
Distinct (%)	100.0%
Missing	10
Missing (%)	26.3%
Infinite	0
Infinite (%)	0.0%
Mean	1.5613566
Minimum	0.96106456
Maximum	2.5882353
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	432.0 B



Quantile statistics

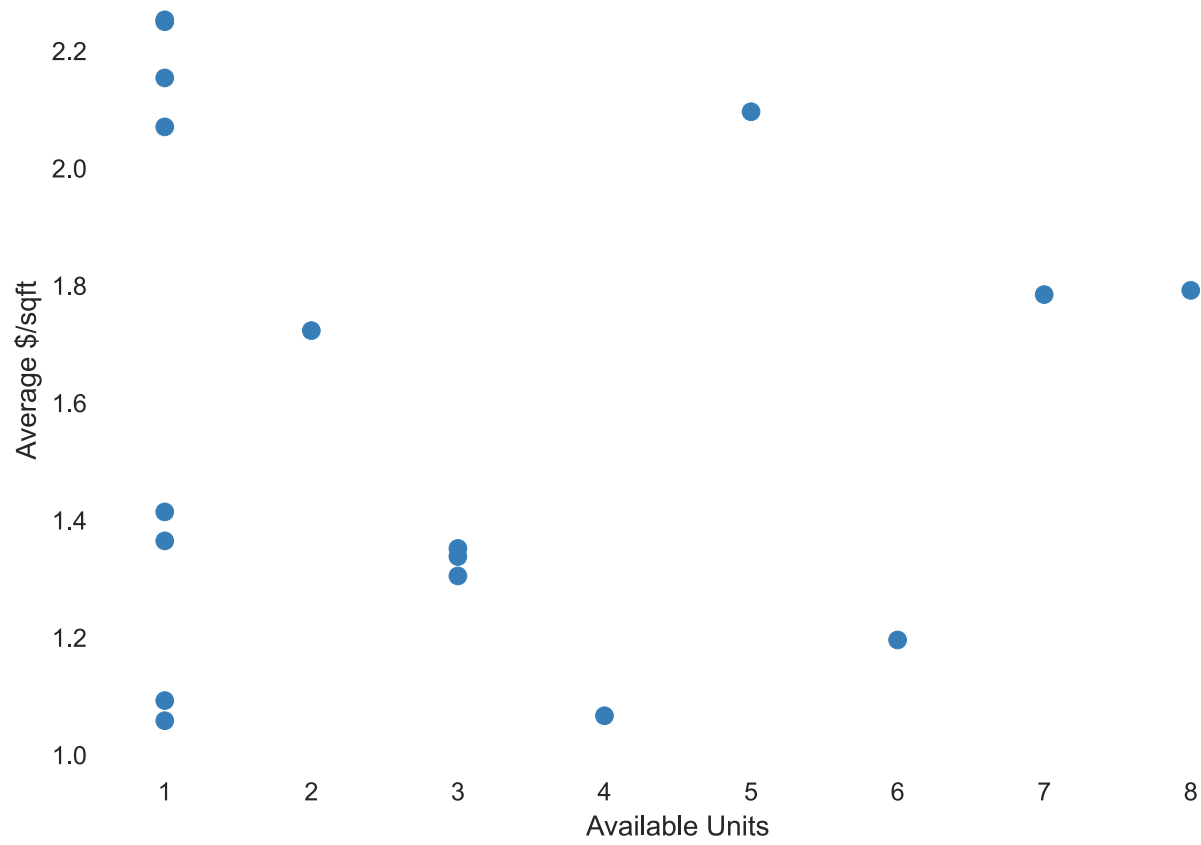
Minimum	0.96106456
5-th percentile	1.0224989
Q1	1.2436019
median	1.3845211

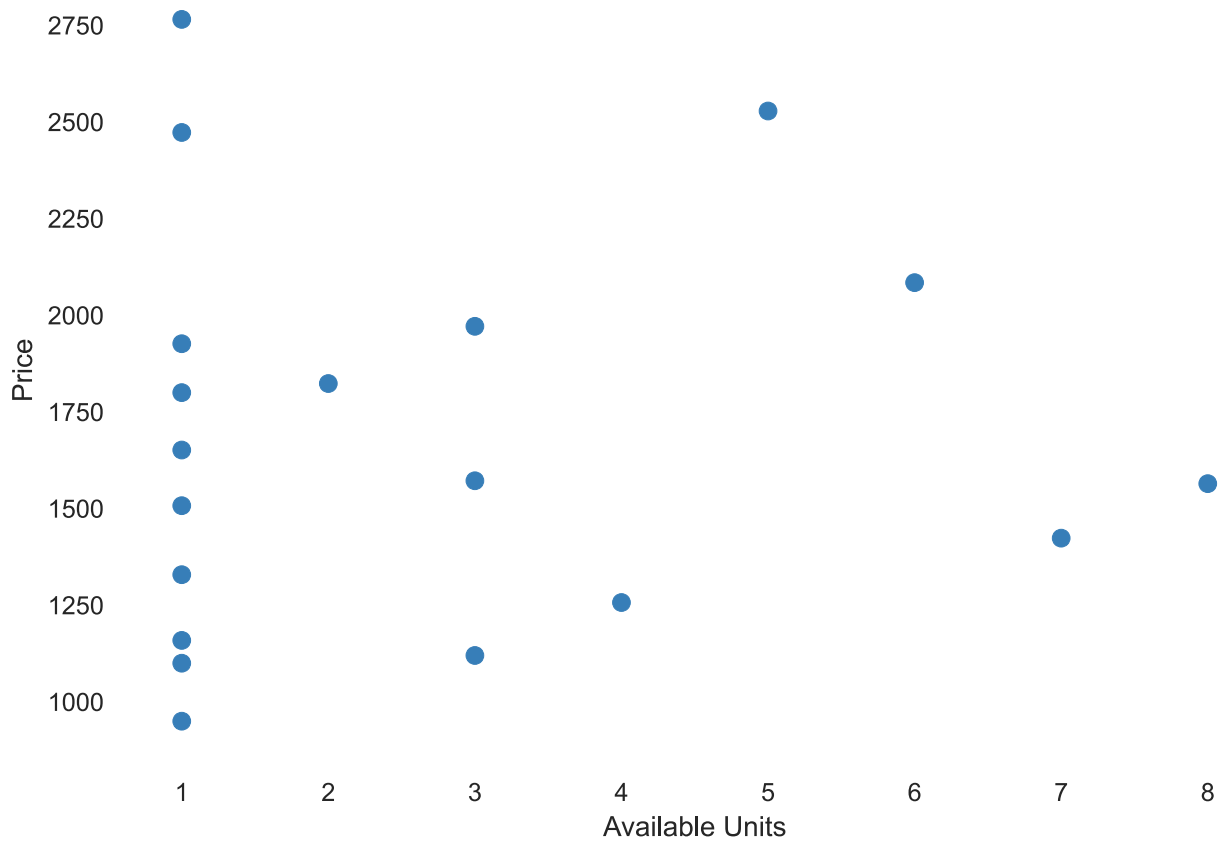
Q3	1.8617622
95-th percentile	2.3872966
Maximum	2.5882353
Range	1.6271707
Interquartile range (IQR)	0.61816023

Descriptive statistics

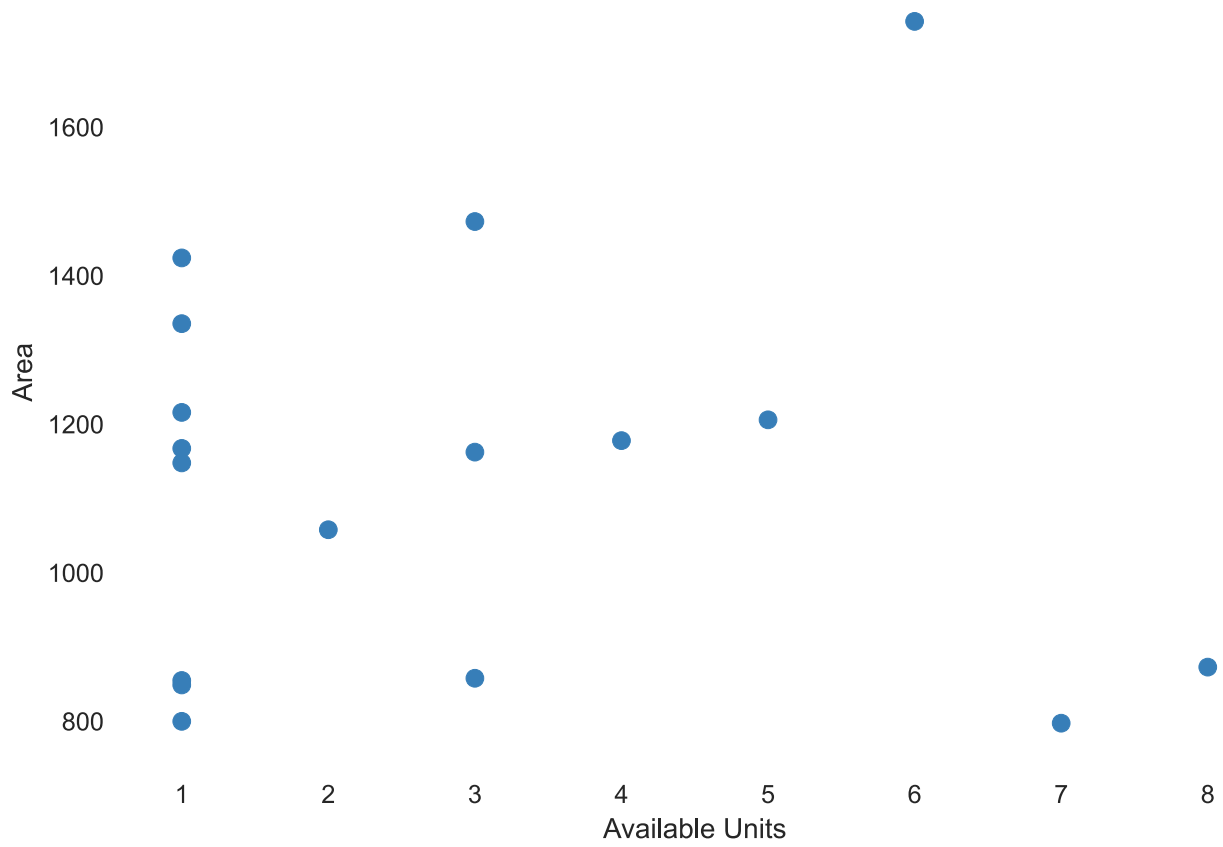
Standard deviation	0.47445548
Coefficient of variation (CV)	0.30387388
Kurtosis	-0.6300116
Mean	1.5613566
Median Absolute Deviation (MAD)	0.30946963
Skewness	0.73489324
Sum	43.717984
Variance	0.22510801
Monotonicity	Not monotonic

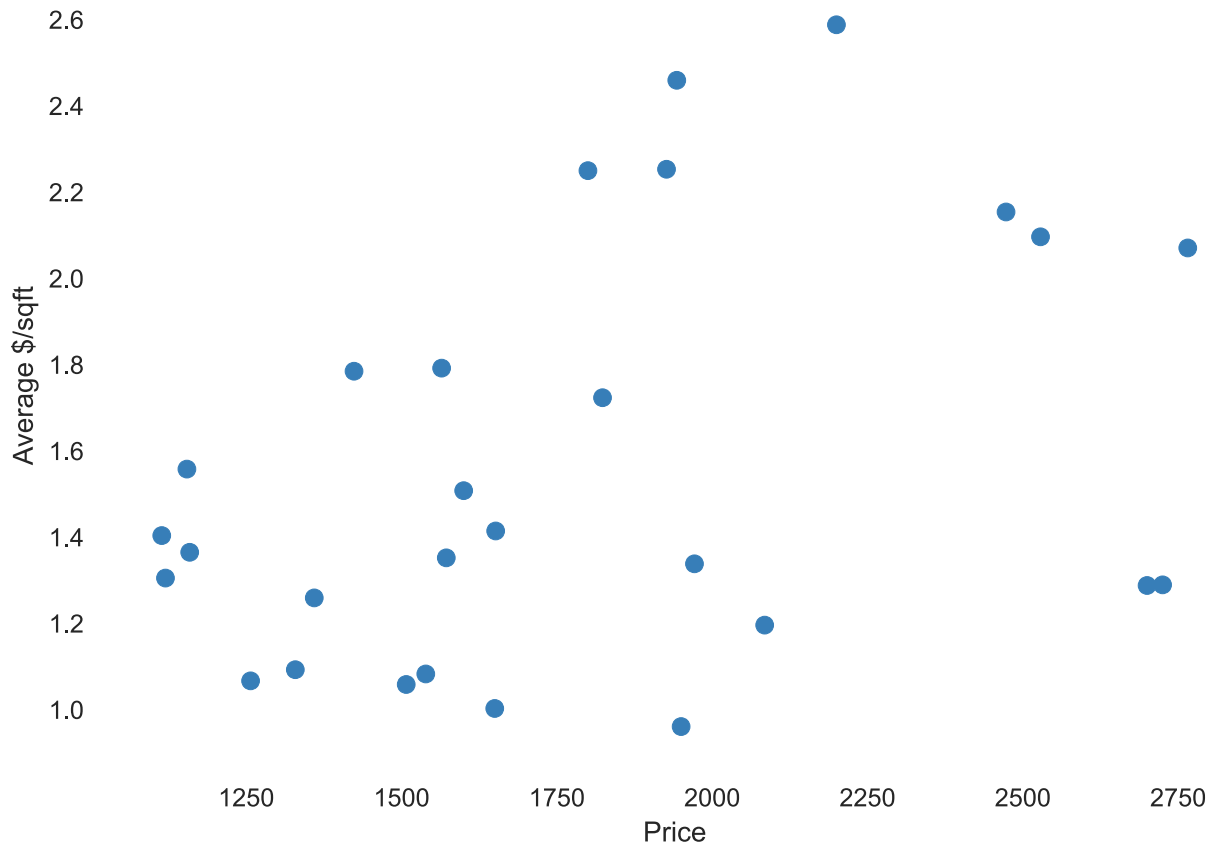
# Interactions

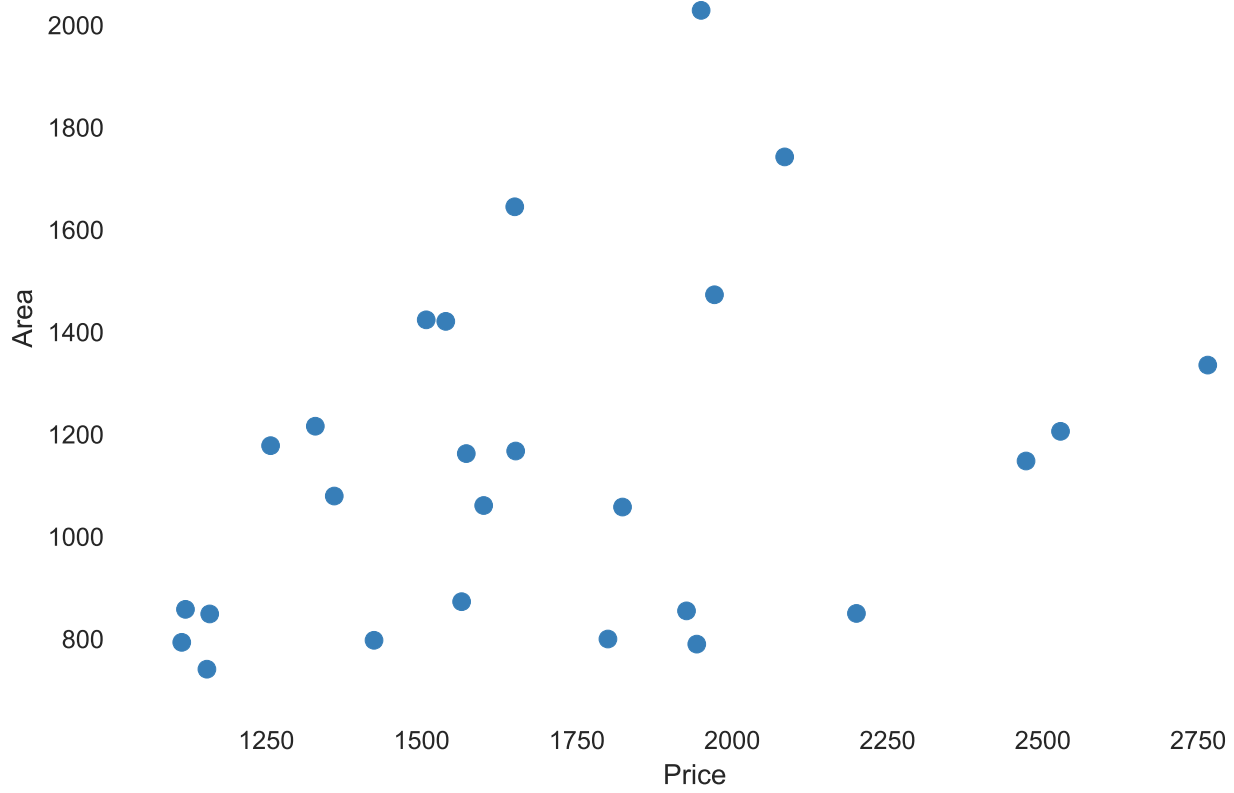


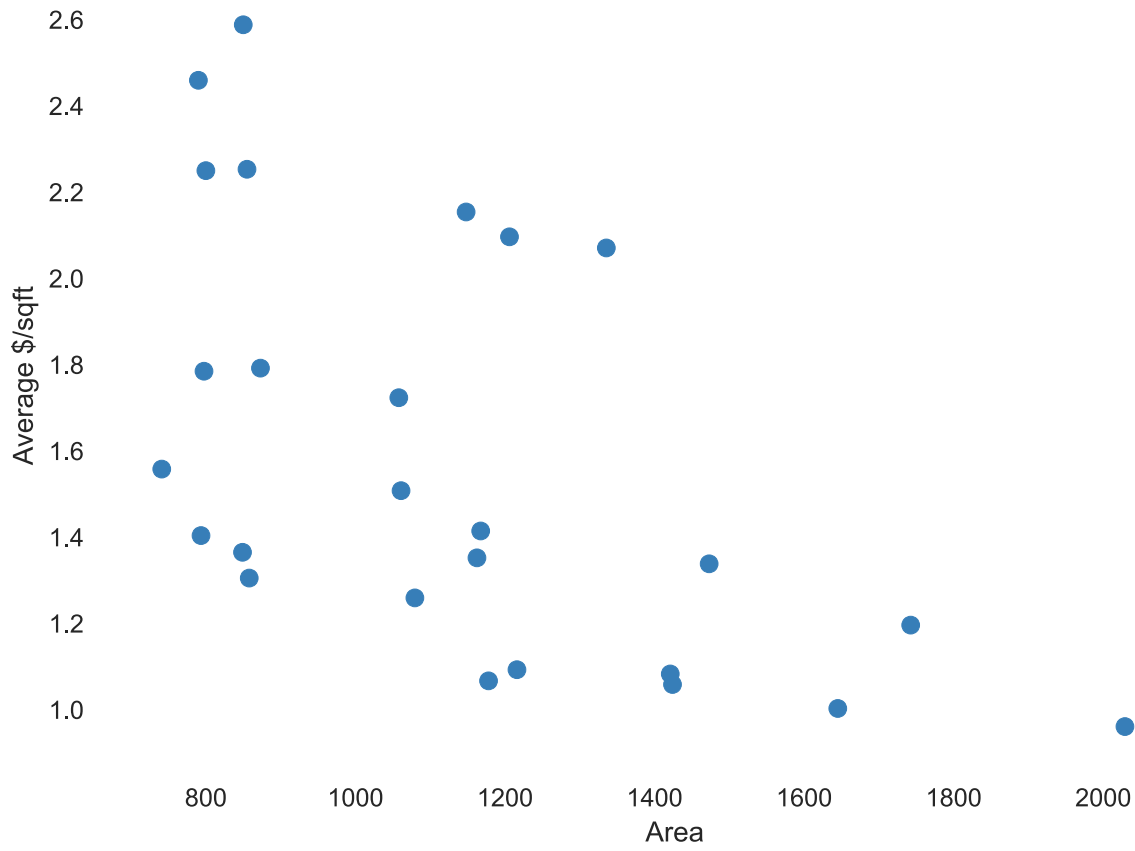




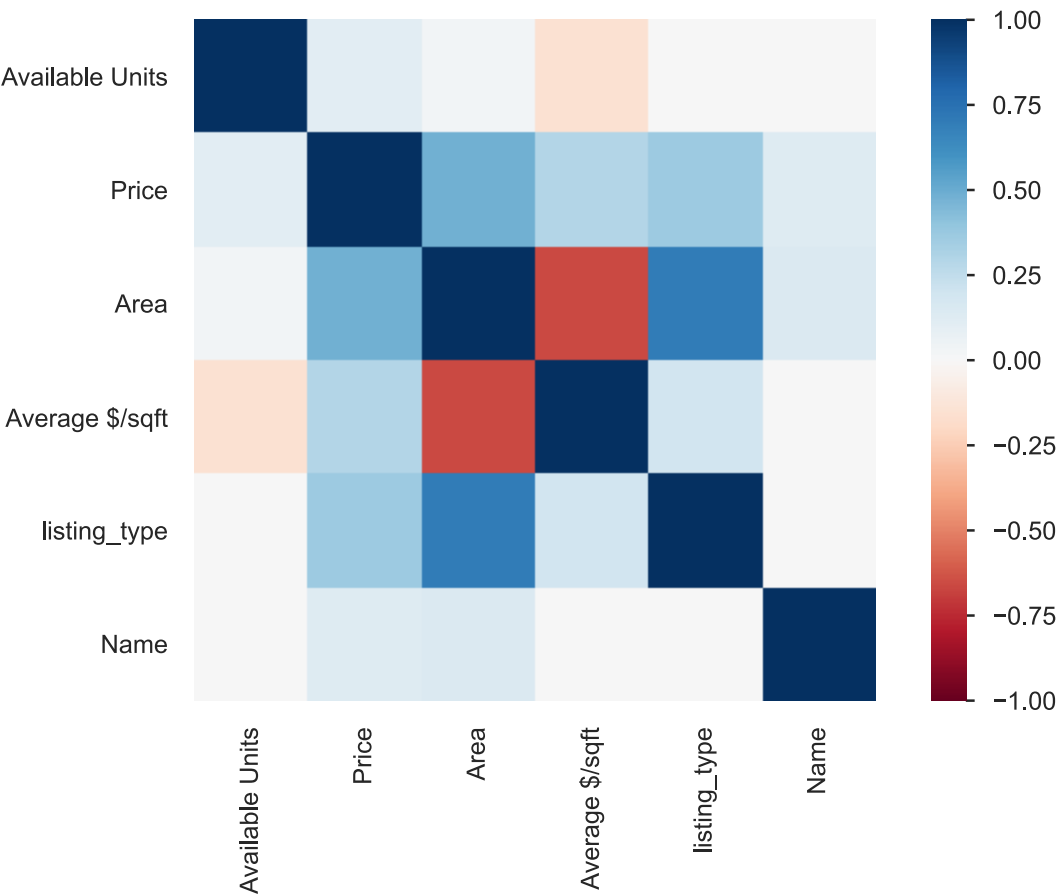






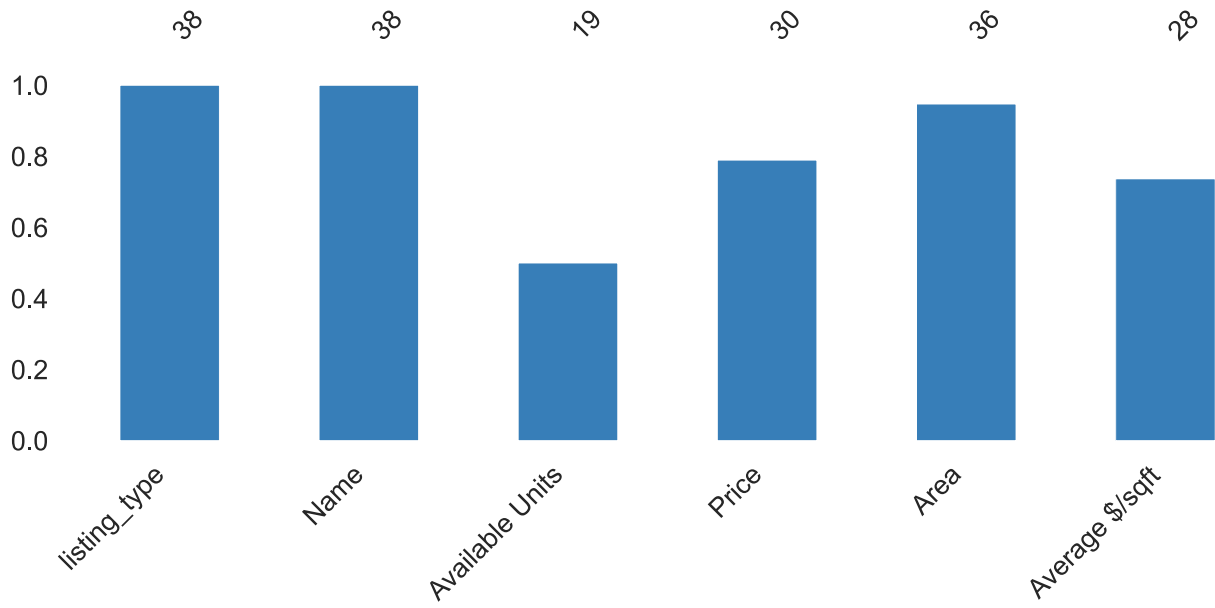


# Correlations

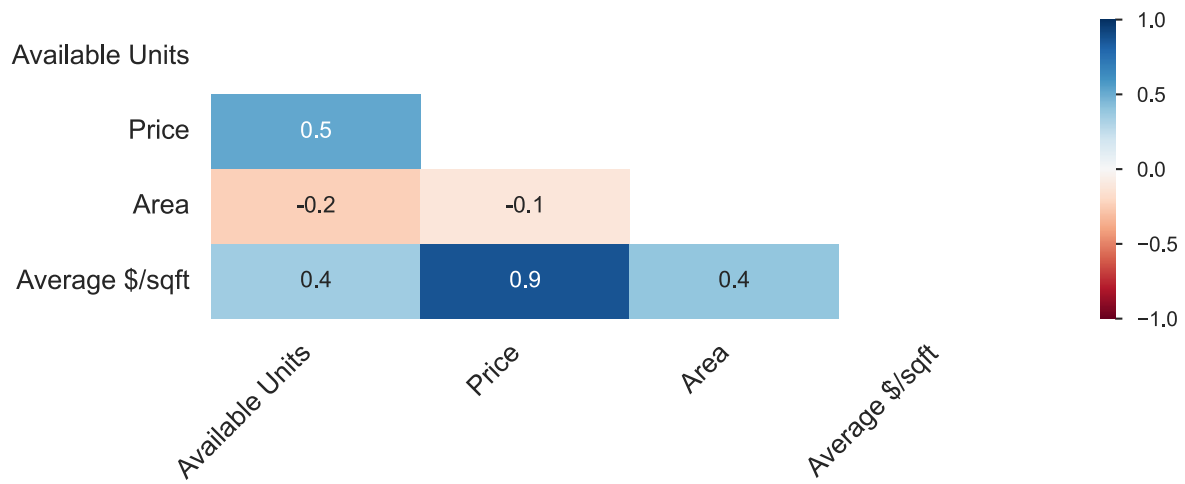


	Available Units	Price	Area	Average \$/sqft	listing_type	Name
Available Units	1.000	0.106	0.027	-0.151	0.000	0.000
Price	0.106	1.000	0.482	0.290	0.361	0.126
Area	0.027	0.482	1.000	-0.662	0.695	0.143
Average \$/sqft	-0.151	0.290	-0.662	1.000	0.193	0.000
listing_type	0.000	0.361	0.695	0.193	1.000	0.000
Name	0.000	0.126	0.143	0.000	0.000	1.000

# Missing values



A simple visualization of nullity by column.



The correlation heatmap measures nullity correlation: how strongly the presence or absence of one variable affects the presence of another.



# Sample

listing_type		Name	Available Units	Price	Area	Average \$/sqft
0	1	Madison Gardens	NaN	NaN	850.0	NaN
1	1	Main Street Apartments	1.0	1926.5	855.0	2.253216
2	1	Laurel at Dry Creek	8.0	1564.5	873.0	1.792096
3	1	Mosby at Bridgestreet	7.0	1423.5	797.5	1.784953
4	1	Willow Run Apartments	NaN	NaN	540.0	NaN
5	1	Highland Pointe	1.0	1159.0	849.0	1.365135
6	1	Brixworth at Bridge Street	NaN	1154.5	741.0	1.558030
7	1	201 Pumprock Dr	1.0	1100.0	NaN	NaN
8	1	Emerald Ridge	NaN	1114.0	793.5	1.403907
9	1	Royal Pines	NaN	NaN	750.0	NaN

**As for the current rental price, although the report shows that there are a considerable percentage of missing data, these missing values won't be taken into account since we should only consider the rental properties that are available for us to lease.**

Nevertheless, there are only 38 available data observations found on rent.com, the source data is too small to build any complicated model. Plus building a regression model based on data with such quality will definitely produce an expensive error. Considering the limited amount of raw data obtained and the poor quality of these data, the estimation of rental price in current market will be simply computed based on descriptive analysis.

```
#Rental Price for 1-bed rental properties in zip code 35806 area
Pr=[0,0,0]
Un=[0,0,0]
for index, r in bed.iterrows():
    if r['listing_type']==1:
        if not math.isnan(r['Available Units']) and not math.isnan(r['Average $/sqft']):
            Pr[0]+=r['Average $/sqft']*r['Available Units']
            Un[0]+=r['Available Units']
    if r['listing_type']==2:
        if not math.isnan(r['Available Units']) and not math.isnan(r['Average $/sqft']):
            Pr[1]+=r['Average $/sqft']*r['Available Units']
            Un[1]+=r['Available Units']
    if r['listing_type']==3:
        if not math.isnan(r['Available Units']) and not math.isnan(r['Average $/sqft']):
            Pr[2]+=r['Average $/sqft']*r['Available Units']
            Un[2]+=r['Available Units']
RP_average=[i/j for i,j in zip(Pr,Un)]
```

RP\_average

[1.7436131595082263, 1.3694804378556216, 1.550330789206396]

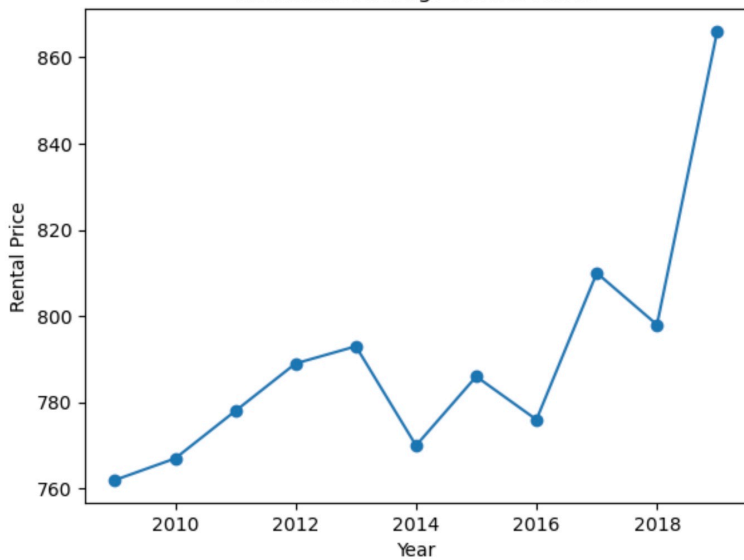
# Next, prediction for the future markets

As the result above shows, the current rental prices for 1-bed, 2-bed, 3-bed rental properties taking average of the available data in the market are 1.744\$/sqft, 1.369\$/sqft, and 1.550\$/sqft respectively.

Again, the data provided by rent.com is lack of quality and quantity. It's impossible to build contextual time-based analysis using data like that. To solve this problem, I searched for rent data over the past few years in the greater Huntsville area. I'll manage to use these data to generate a time series forecasting model and use this model to predict the rental price in zip code 35806.

**(Data Source:<https://www.deptofnumbers.com/rent/alabama/huntsville/>)**

Huntsville Average Rental Prices



**By the ARIMA model's prediction, the average housing price in Huntsville in 2025 will be 764.283483, decreasing from 765.003955 in 2023.**

Let's assume that the rental price in zip code 35806 will fluctuate along with the average housing price in the greater Huntsville area.

# Therefore, the resulted prediction for the rental price in zip code 35806 area in 2025 will be 1.742\$/sqft, 1.368\$/sqft, and 1.549\$/sqft respectively for 1-bed, 2-bed, and 3-bed rental properties.