REAL ESTATE TRENDS & INVESTIGATING RELATIONSHIPS WITH COVID-19

Data Science Capstone Project
Exploratory Data Analytics Report

Date:

[11/23/2020]

Team Members: 4

Name: Lawrence Love

Name: Gustavo Ferreira

Name: Frank Zhao

Name: Yan Li

Analysis the basic metrics of variables

House for Sale

Discrete Variables: baths_full, baths, beds, photo_count, baths_half, agent_id

Continuous Variables: price, lat, lon, building_size(sqft), lot_size(sqft)

Categorical Variables: property_id, prop_type, prop_sub_type, prop_status, city, line, state_code, county, neighborhood_name, agent_name, brand_name, postal_code

Size: 9562 instances, 23 Features

Statistics of the Price Variable:

Mean	Standard Deviation	Min	Max	25% Percentiles	50% Percentiles	75% Percentiles
381703.56	559604.07	6000	25000000	156250	265000	434900

House for Rent

Discrete Variables: year_built, beds, baths_full, baths, photo_count, garage

Continuous Variables: price, lat, lon, building_size(sqft), lot_size(sqft)

Categorical Variables: property_id, prop_type, list_date, last_update, city, line, state_code, county, neighborhood_name, status, brand_name, broker_name, postal_code

Size: 5277 instances, 25 Features

Statistics of the Price Variable:

Mean	Standard Deviation	Min	Max	25% Percentiles	50% Percentiles	75% Percentiles
1808.65	881.32	334	12000	1300	1625	2070

Sold Houses

Discrete Variables: total_homes_sold, median_days_to_close, total_new_listings, average_new_listings, inventory, total_active_listing, age_of_inventory, median_days_on_market

Continuous Variables: median_sale_price, price_drops, percent_active_listings_with_price_drops, pending_sales, median_new_listing_price, homes_delisted, median_active_list_price,avg_offer_to_list, months_of_supply, percent_total_price_drops_of_inventory

Categorical Variables: period_begin, period_end, duration

Size: 200 instances, 21 Features

Statistics of the median_sale_price Variable:

Mean	Standard Deviation	Min	Max	25% Percentiles	50% Percentiles	75% Percentiles
206690.14	24408.54	143000	260000	189800	204950	220062.5

COVID-19 Cases by zip code – There were 14 different COVID-19 datasets used in this project. This is the main dataset of focus is below:

Discrete Variable: zip_code

Categorical Variable: covid_status

Continuous Variable: count

Size: 116 instances, 3 features

Statistics of count variable:

Mean	Standard Deviation	Min	Max	25% Percentiles	50% Percentiles	75% Percentiles
6946.81	8902.00	143000	34922	370.50	1775.00	12522.00

Non-graphical and graphical univariate analysis

House for Sale

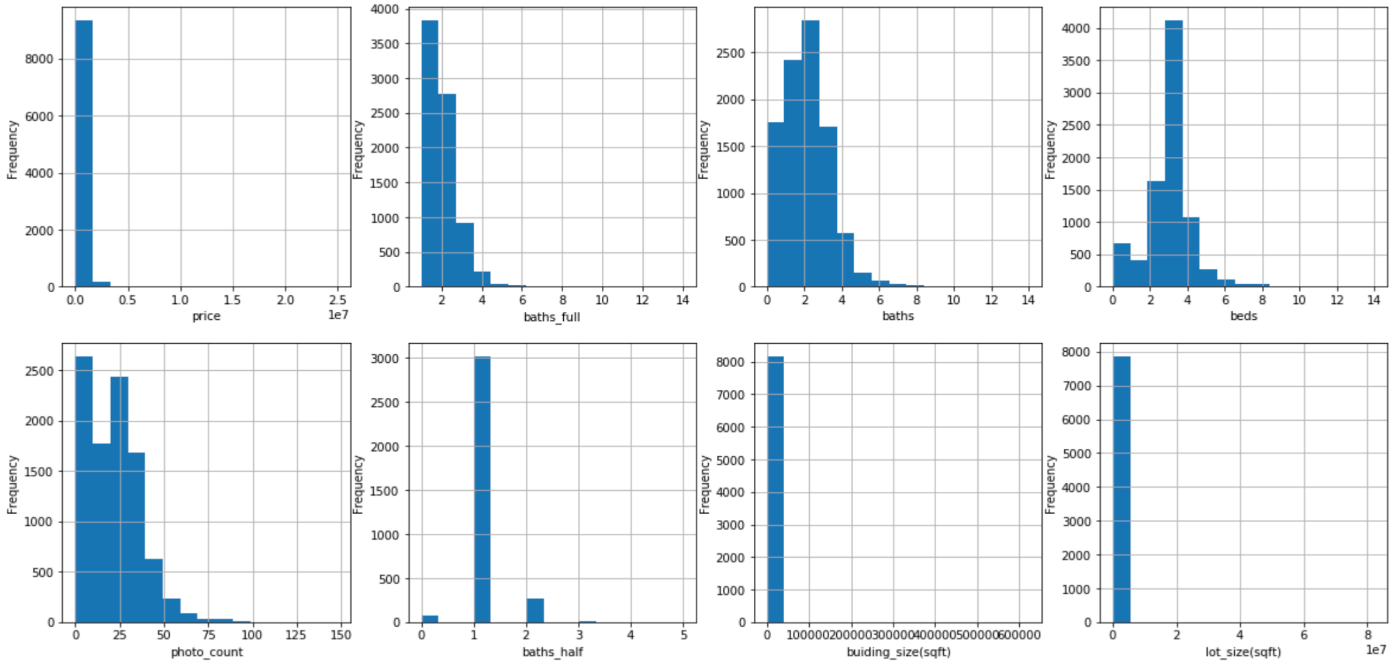

Figure 1 Distribution of price, baths_full, baths, beds, photo_count, baths_half, building_size(sqft), lot_size(sqft)

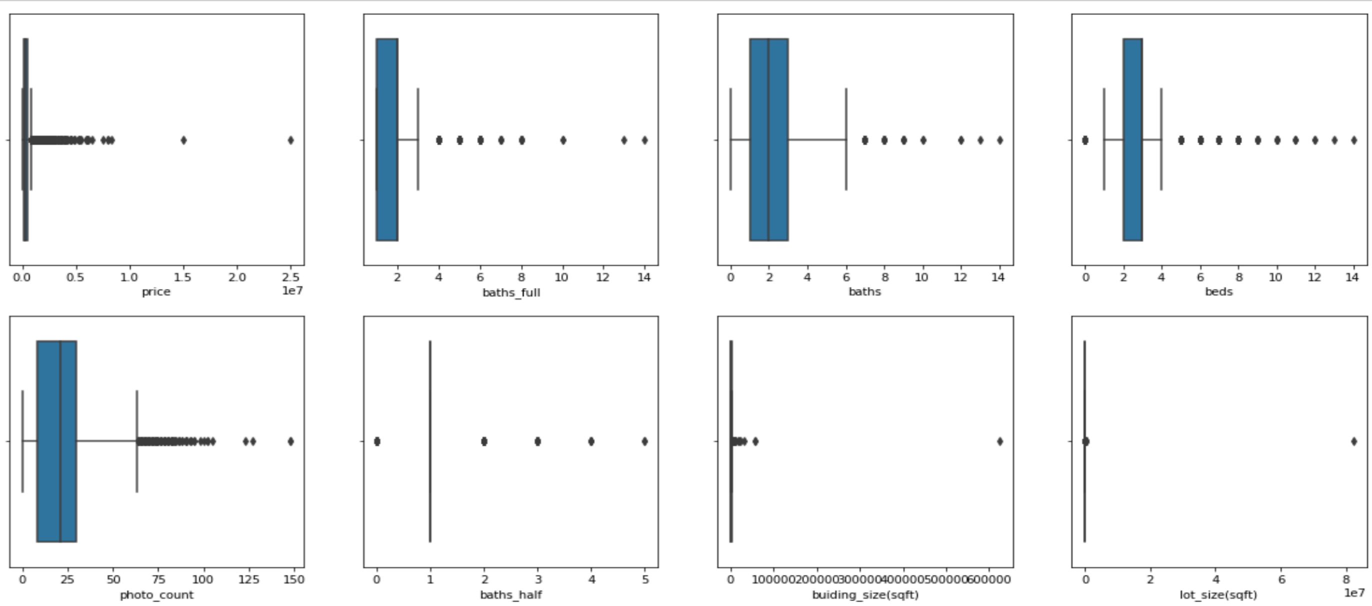

Figure 2 Boxlpots of *price, baths_full, baths, beds, photo_count, baths_half, building_size(sqft), lot_size(sqft)*

prop_type	counts
condo	7078
land	1182
multi_family	720
single_family	582

Figure 3 Unique Values of prop_type

prop_sub_type	counts
townhomes	5072
condos	1186
duplex_triplex	792

Figure 4 Unique Values of prop_sub_type

neighborhood_name	counts
Center City	1728
South Philadelphia	833
West Philadelphia	609
Lower North	458
Far Northeast Philadelphia	356
Kensington	353
Upper North Philadelphia	353
North Delaware	329
Near Northeast Philadelphia	287
Point Breeze	287

Figure 5 TOP 10 Unique Values of neighborhood_name

postal_code	counts
19146	636
19147	512
19148	468
19121	461
19123	397
19125	383
19122	368
19103	367
19145	366
19134	347

Figure 6 TOP 10 Unique values of postal_code

House for Rent

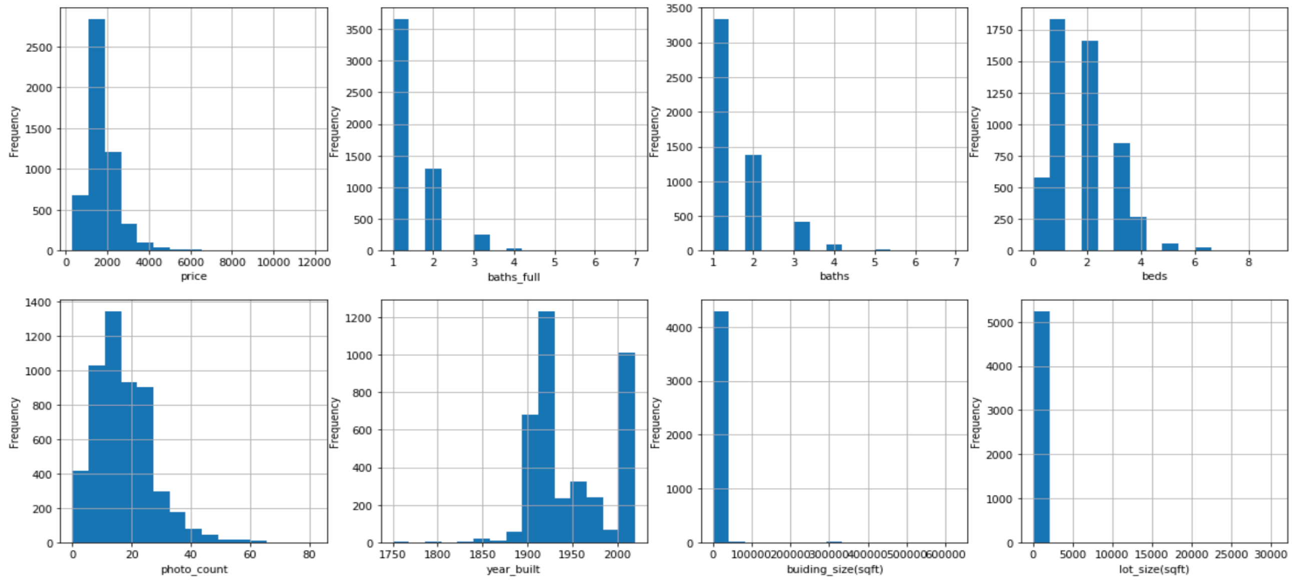

Figure 7 Distribution of price, baths_full, baths, beds, photo_count, year_built, building_size(sqft), lot_size(sqft)

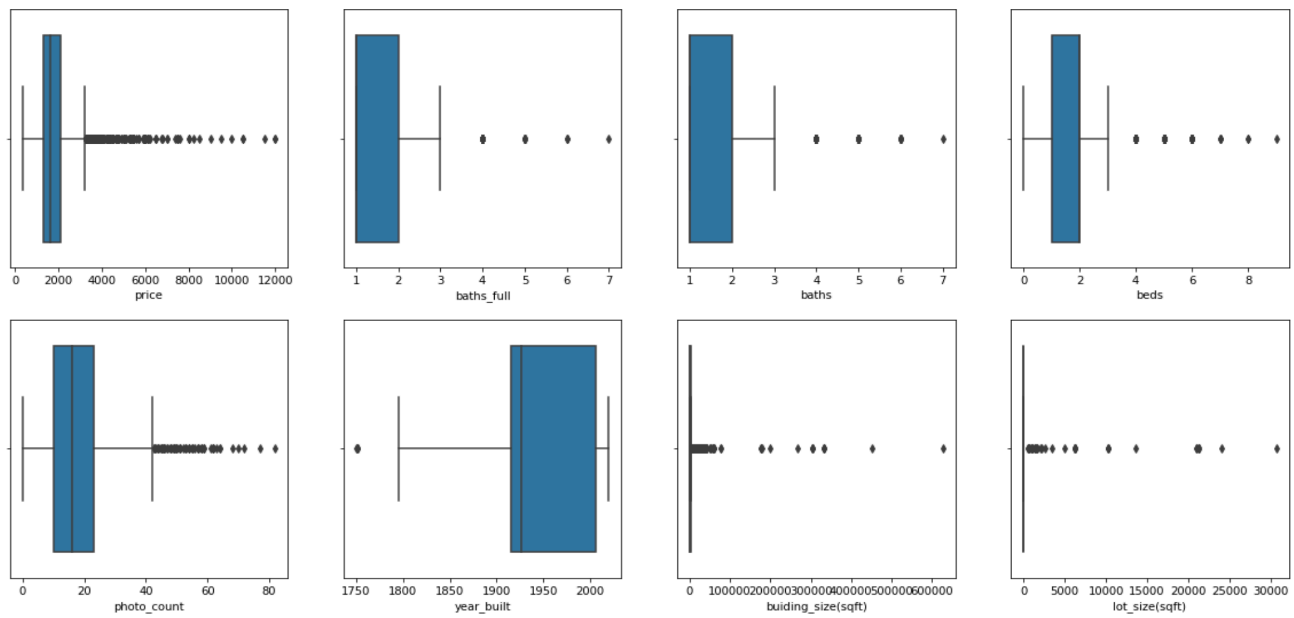

Figure 8 Boxplots of price, baths_full, baths, beds, photo_count, year_built, building_size(sqft), lot_size(sqft)

prop_type	counts
condo	3217
townhome	1116
apartment	638
single_family	214
duplex_triplex	86
other	4
multi_family	2

Figure 9 Unique Values of prop_type

postal_code	counts
19103	800
19107	469
19106	389
19146	356
19121	346
19102	343
19147	334
19130	308
19123	272
19122	234

Figure 10 TOP 10 Unique values of postal_code

	counts
neighborhood_name	
Center City	1231
Rittenhouse	580
Logan Square	367
North Central	202
Fishtown	128
Graduate Hospital	124
Point Breeze	120
Rittenhouse Square	101
Queen Village	97
East Kensington	96

Figure 11 TOP 10 Unique Values of neighborhood_name

Sold Houses

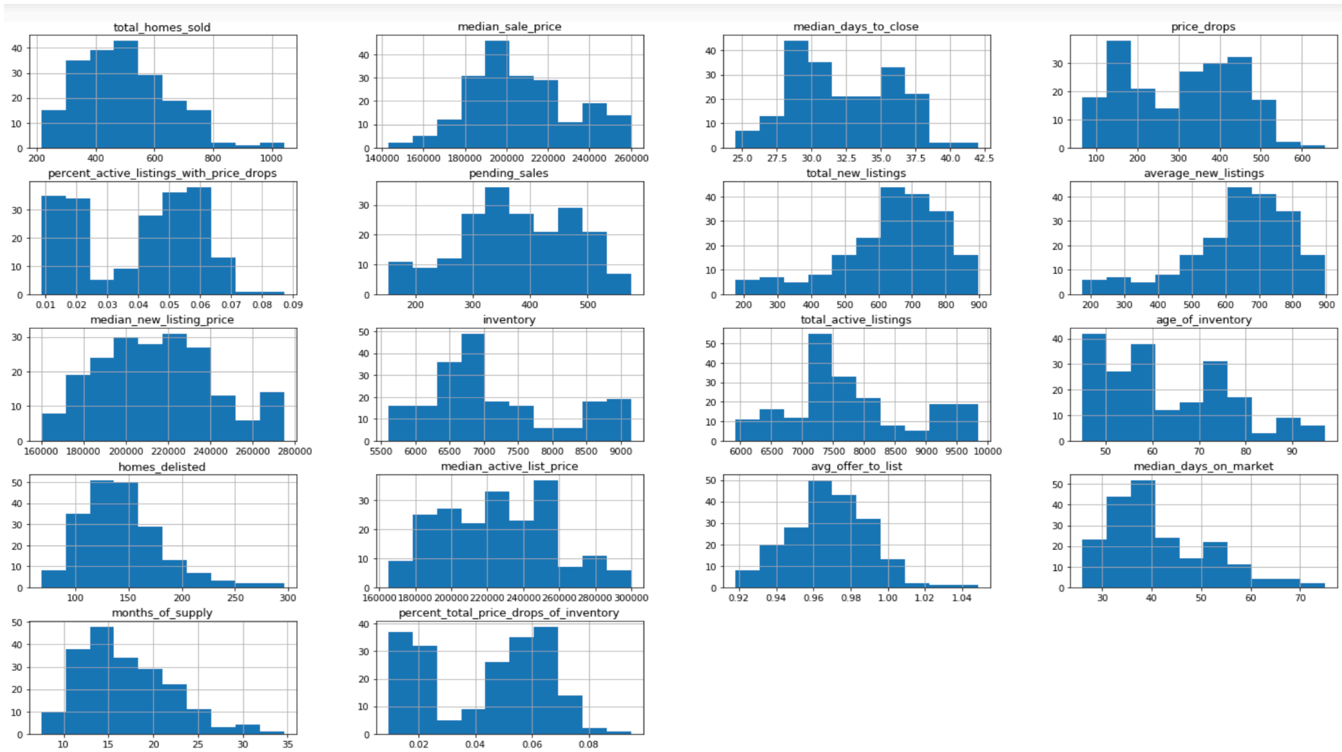

Figure 12 Distribution of All the Numerical Variables

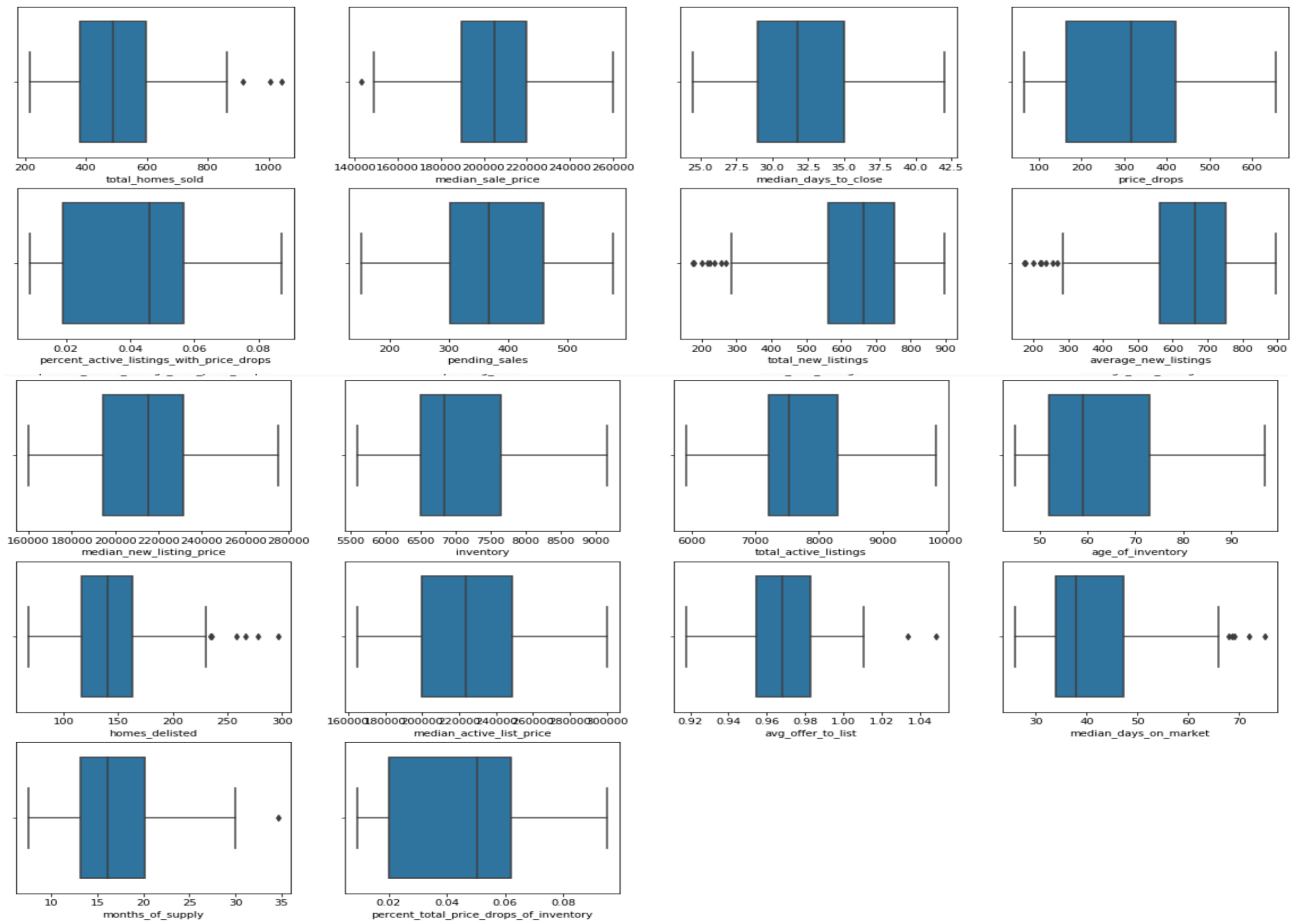

Figure 13 Boxplots of All the Numerical Variables

COVID-19 Cases by zip code

	covid_status	zip_code	count
0	NEG	19140	21439
1	POS	19127	156
2	NEG	19133	9398
3	POS	19146	1266
4	NEG	19138	13570
5	NEG	19152	19149
6	NEG	19188	6
7	POS	19115	1672
8	NEG	19144	21584
9	NEG	19141	12969

Figure 14 First ten rows of dataset

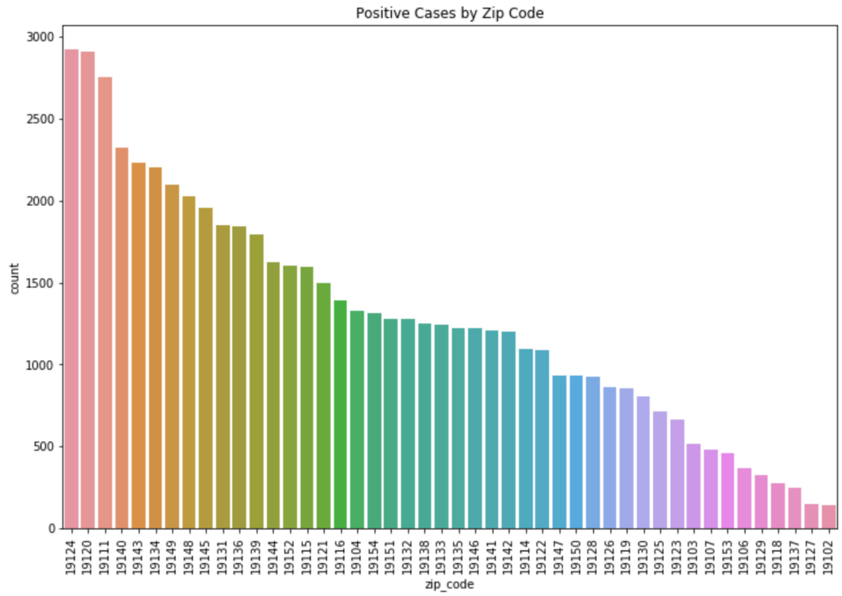

Figure 15 Barplot for positive cases by zip code

Missing value analysis and outlier analysis

Missing value

In our "house for sale" dataset, there are 9562 instances and 31 features after the first requesting, and 9562 instances and 18 features after removing unuseful ones. Among those 18 features, 7 features are containing missing values. As shown below:

property_id	0	beds	1182
prop_type	0	building_size	1381
prop_sub_type	2512	agents	94
address	0	last_update	0
branding	0	photo_count	0
prop_status	0	page_no	0
price	0	rank	0
baths_full	1760	lot_size	1593
baths	0	baths_half	6168

Figure 16 Missing Values

prop_sub_type: Since we have "prop_type" without missing values, we don't have to worry about it, and we can drop it.

Baths_full, beds, building_size, lot_size, and baths_half: For these numerical features, we fill up the missing value by 0. It's easy to understand that if a property has that missing values in those features, we can assume it doesn't have any in those features.

Agents: Since we don't include this feature in visualization, we just don't change anything to it. However, for our second stage, the model training, we may use this feature and by that time, we will remove the rows that have missing values in this feature. It won't affect that much by removal since there are only 94 missing values.

Historical and Current Listing Data

In conducting statistical analysis on the historical and current listing datasets, we used various tests for equal variance to help determine that we did not have equal variance. According to boxplots, both datasets had outliers for price ranging from ~$1,000,000 up to ~$25,000,000. Based on the visual provided by the boxplot, it was determined to consider all data with prices greater than $899,999 outliers. After re-running the variance tests, we now had results to suggest equal variance between the datasets.

Feature engineering and analysis

This project contained no predictive modeling or machine learning. However, some feature engineering took place by creating two different "average price per sqft" variables and two different "price per square feet" variables. For these variables, we had data on "lot_size (sqft)" to measure the size of a lot, and "building_size (sqft)" to measure the space inside a building. We then took the price of each listing and divided it by that listing's respective "X_size (sqft)" to get "building_price_per_sqft" and "lot_price_per_sqft".

Additionally, the dataset contained data for postal codes. This allowed us to analyze regional data within Philadelphia. Using the postal codes, we were then able to use our new "price_per_sqft" to create and analyze data for the average price per sqft for both the lots and the buildings in each postal code.

We found that some properties have a value of zero in either "building_size" or "lot_size" which lead to a result of "inf" after applying division to get "building_price_per_sqft" and "lot_price_per_sqft". Thus, we remove properties with 0 in either "building_size" or "lot_size" to create a more balanced dataset for visualization. Every property will have no zeros in both "building_size" and "lot_size", and when calculating the average unit price in a postal code region, these two features will be counted.

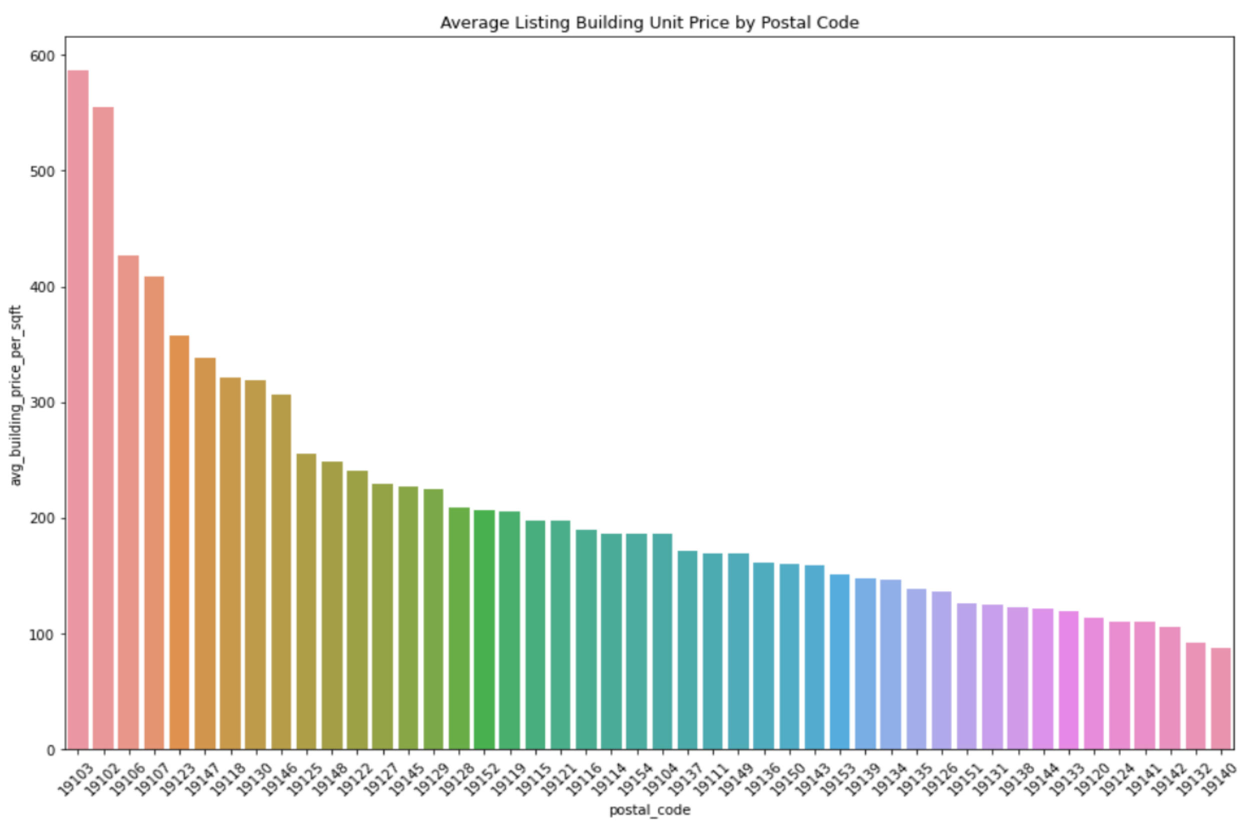

Figure 17 Average building price per sqft by postal code (before removal)

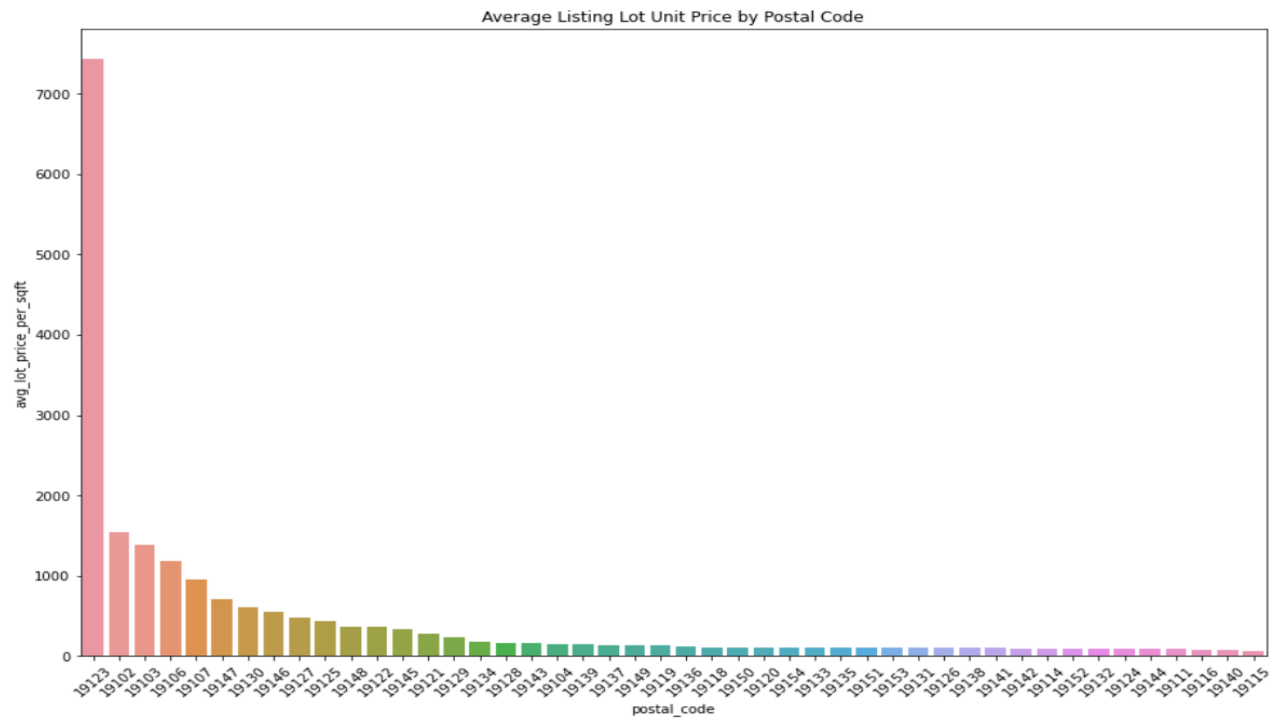

Figure 18 Average lot price per sqft by postal code (before removal)

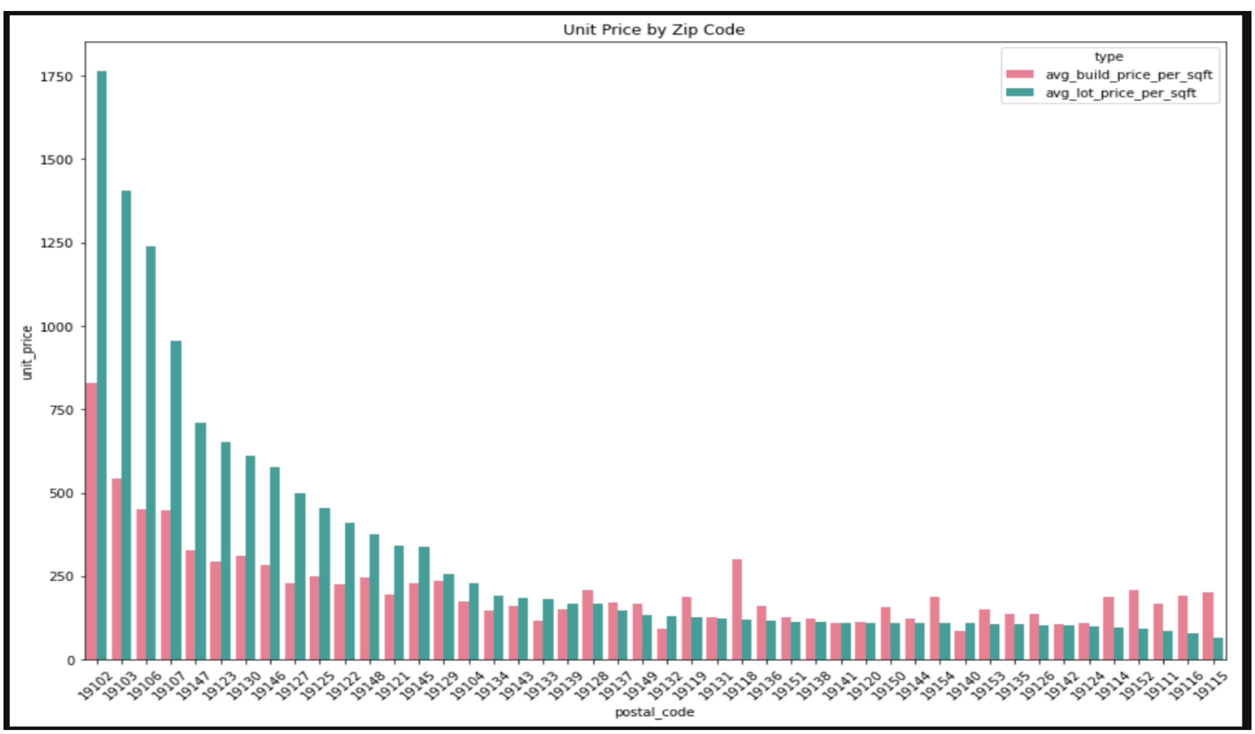

Figure 19 Average building price per sqft by postal code (After removal)

We can see that before removing property with zero in those two features, the region of 19123 shows a significantly high average value in "average lot size", which shows the most expensive region and does not make sense. By applying removal, we now see a more reasonable visualization.

Appendix

1.Code for Figure 1 Distribution of price, baths_full, baths, beds, photo_count, baths_half, building_size(sqft), lot_size(sqft)

```
plt.figure(figsize= (20, 10))
columns = ['price', 'baths_full', 'baths', 'beds', 'photo_count',
           'baths_half',  'buiding_size(sqft)', 'lot_size(sqft)']

for i,col in enumerate(columns):
    plt.subplot(2, 4, i+1)
    sale[col].hist(bins=15)
    plt.xlabel(col)
    plt.ylabel('Frequency')
```

2.Code for Figure 2 Boxplots of price, baths_full, baths, beds, photo_count, baths_half, building_size(sqft), lot_size(sqft)

```
plt.figure(figsize= (20, 10))
columns = ['price', 'baths_full', 'baths', 'beds', 'photo_count',
           'baths_half',  'buiding_size(sqft)', 'lot_size(sqft)']

for i,col in enumerate(columns):
    plt.subplot(2, 4, i+1)
    sns.boxplot(sale[col])
```

Table of Contributions

The table below identifies contributors to various sections of this document.

	Section	Writing	Editing
1	**Analysis the basic metrics of variables**	**L. Love, Y.Li**	**G.Ferreira**
2	**Non-graphical and graphical univariate analysis**	**L. Love, Y.Li, F.Zhao**	**G.Ferreira**
3	**Missing value analysis and outlier analysis**	**L.Love, F.Zhao**	**G.Ferreira**
4	**Feature engineering and analysis**	**L.Love, F.Zhao**	**G.Ferreira**
5	**Appendix**	**Y.Li**	**G.Ferreira**