House Price Prediction in Natural Hazard Prone Areas

Capstone Project 1: EDA & Data Storytelling

Introduction

After collecting data, wrangling data then exploratory analyses were carried out. The following questions were got into my mind and exploratory analyses were done to find answers for all these.

- Geospatial visualization of natural hazard homes in single family and townhouse
- Most common features (bedrooms, bathrooms,year_built, liquefaction, fault zone, landslide, fire hazard) for single family and townhouse
- Distribution of number of houses with price bin for single family and townhouse
- Median number of bedrooms, bathrooms,sqft with price bin for single family and townhouse
- LIquefaction, fault zone,landslide, fire hazard with price bin in single family and townhouse
- Sold price distribution for various zip codes, hazards and non hazards for single family and townhouse
- Best month to put sale for single family and townhouse
- Most popular zip codes saleswise for single family and townhouse
- Number of homes sold from 2016-2019 for single family and townhouse
- Influence of natural hazard on sold price
- Median price/ sqft for hazard and non-hazard homes
- Median price trend over construction year
- Impact of liquefaction,landslides, fault zone, fire hazard on price of single family and townhouse
- People's comment on houses in hazard areas
- Correlation plot between features and price
- Influential features to predict house price

Where are hazards prone areas? Which hazard is most common and least common in city? How are they distributed in the city?

Before going deep into numerical part of data analysis, geospatial analyses were done to observe how hazards are distributed, most common, least common hazards in the city. Folium and basemap packages were used to plot theses maps. Here are the geospatial analyses results for single family and townhouse.

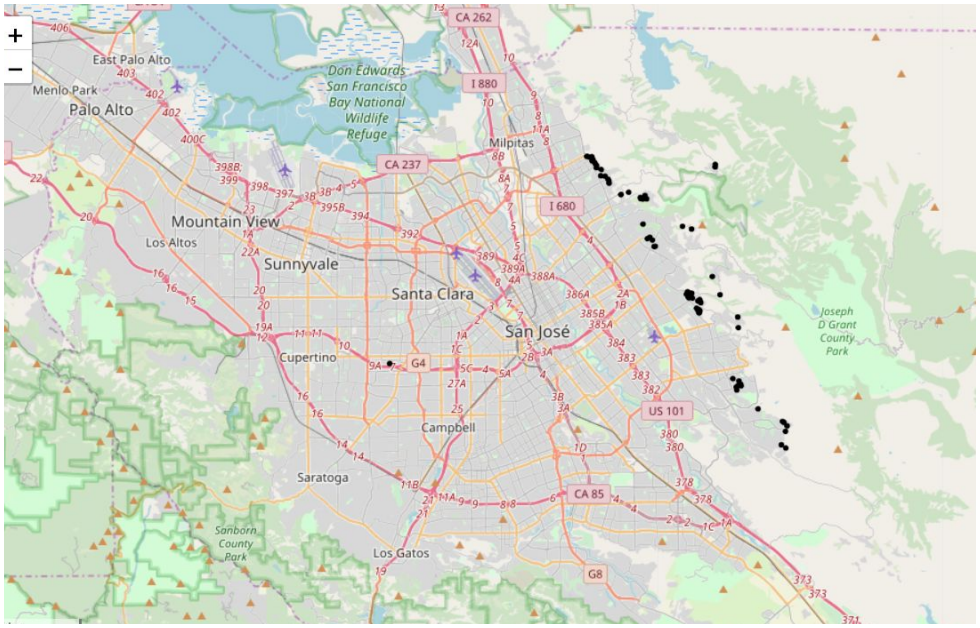


Fig 1a: Fault zone distribution (single family homes)

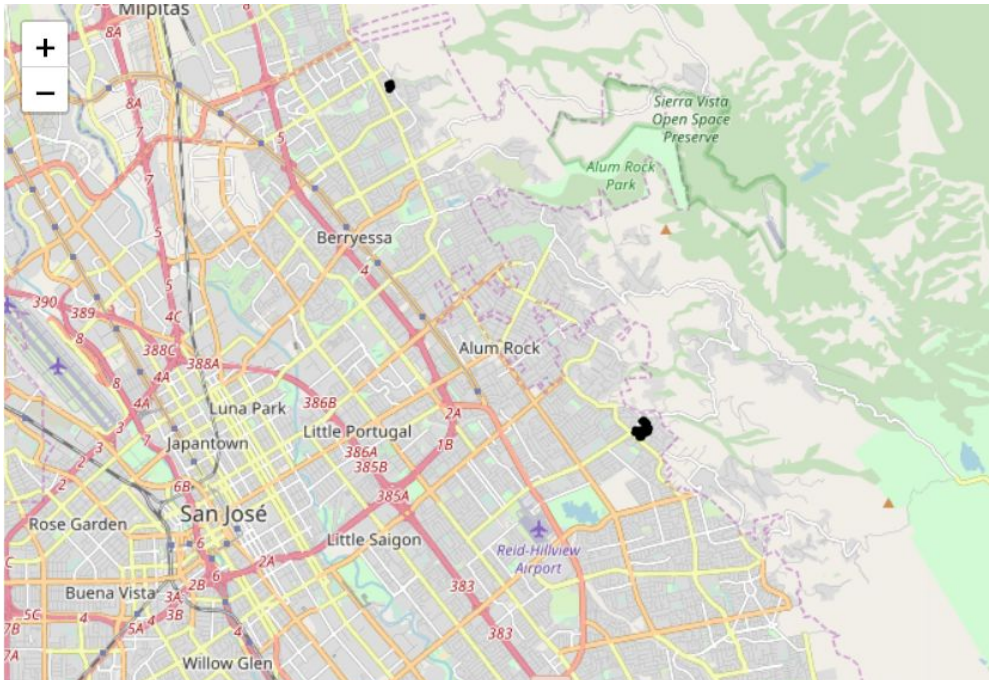


Fig 1b: Fault zone distribution (town homes)

Fault zone distribution points are accumulated in between east san jose land and mountain range (Figs 1a and 1b).

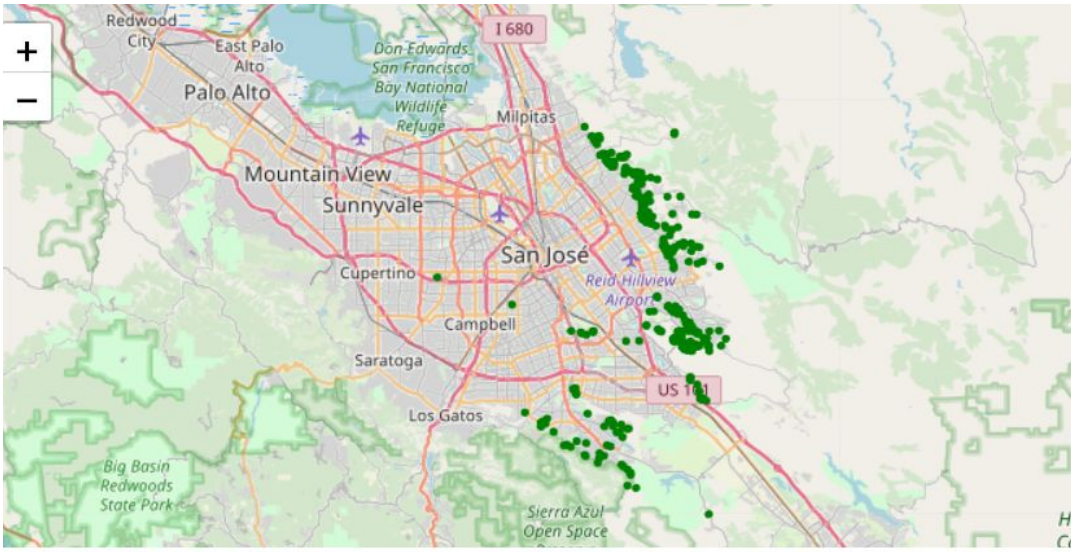


Fig 2a: Landslide distribution (single family homes)

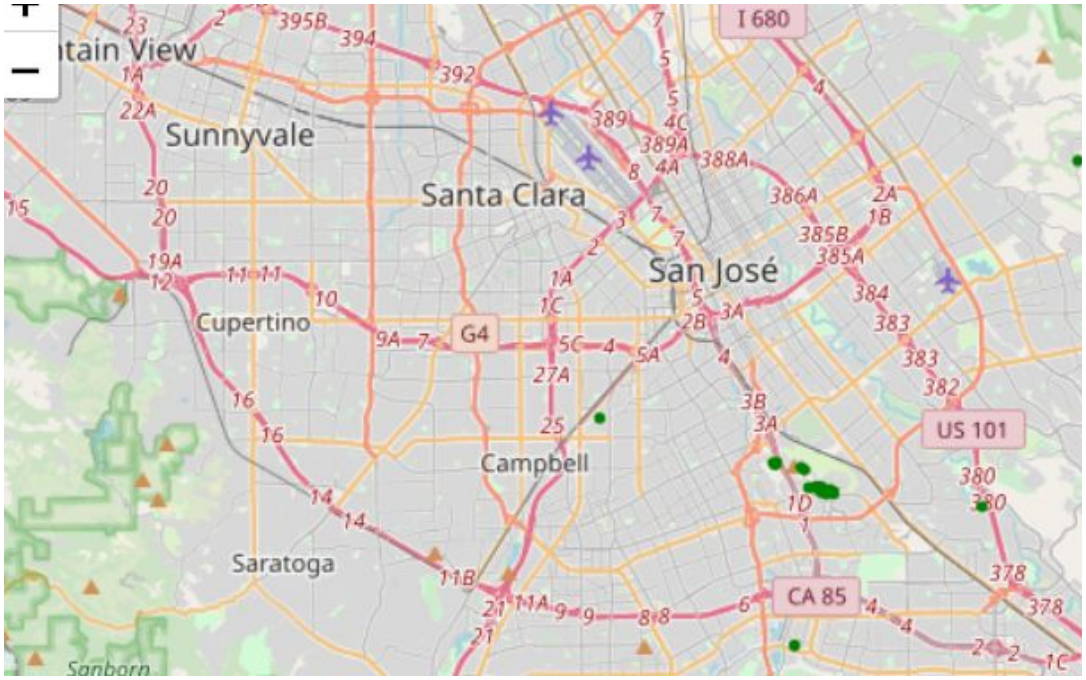


Fig 2b: Landslide distribution (town homes)

Landslide zone distribution points are accumulated along east side mountain range (Figs 2a and 2b).

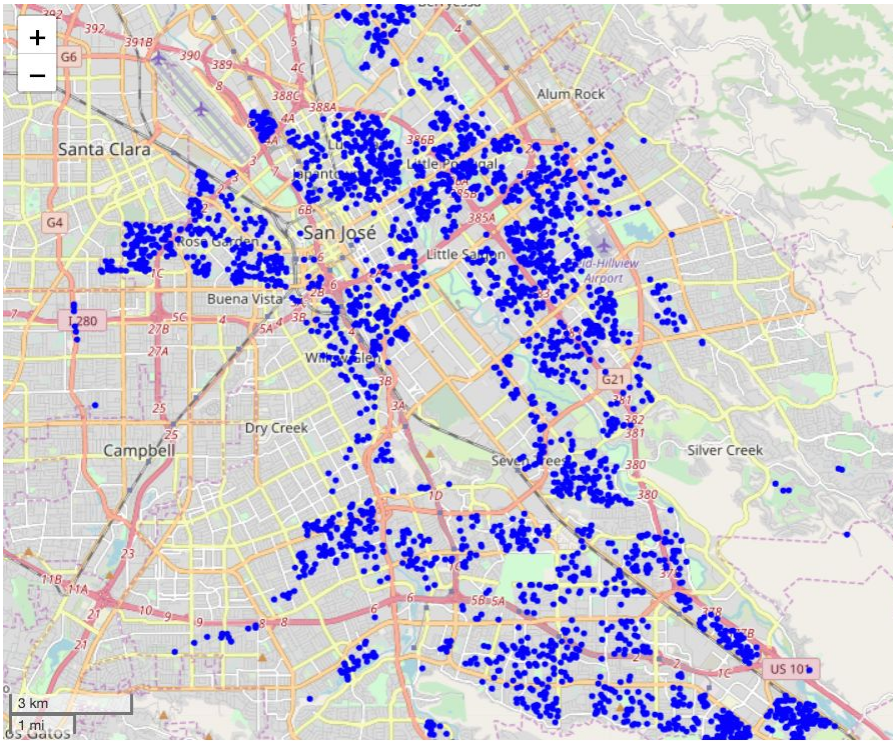


Fig 3a: Liquefaction distribution (single family homes)

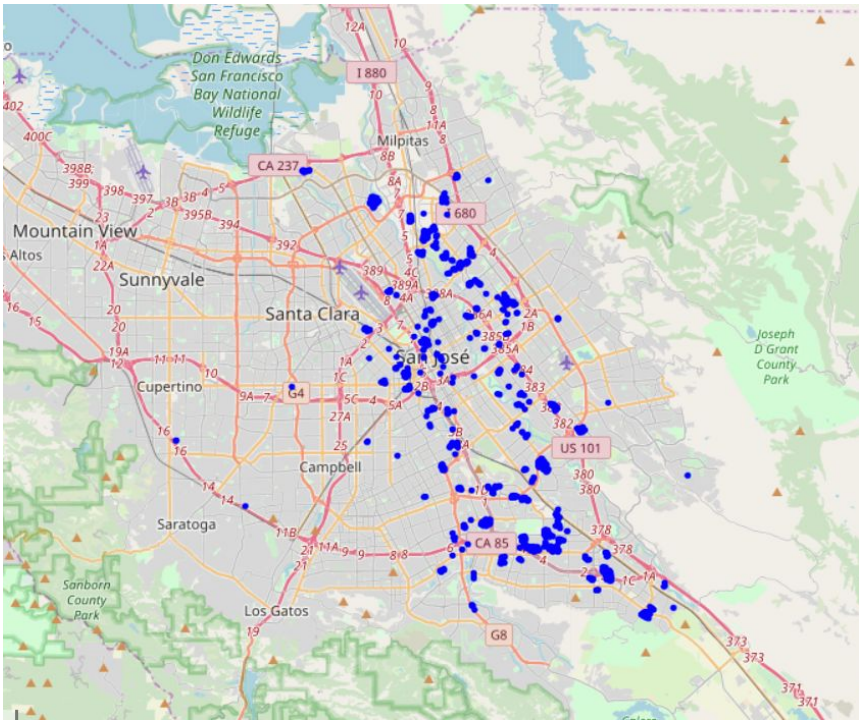


Fig 3b: Liquefaction distribution (town homes)

Liquefaction are the most common hazard and widely distributed in the city (Figs 3a and 3b).

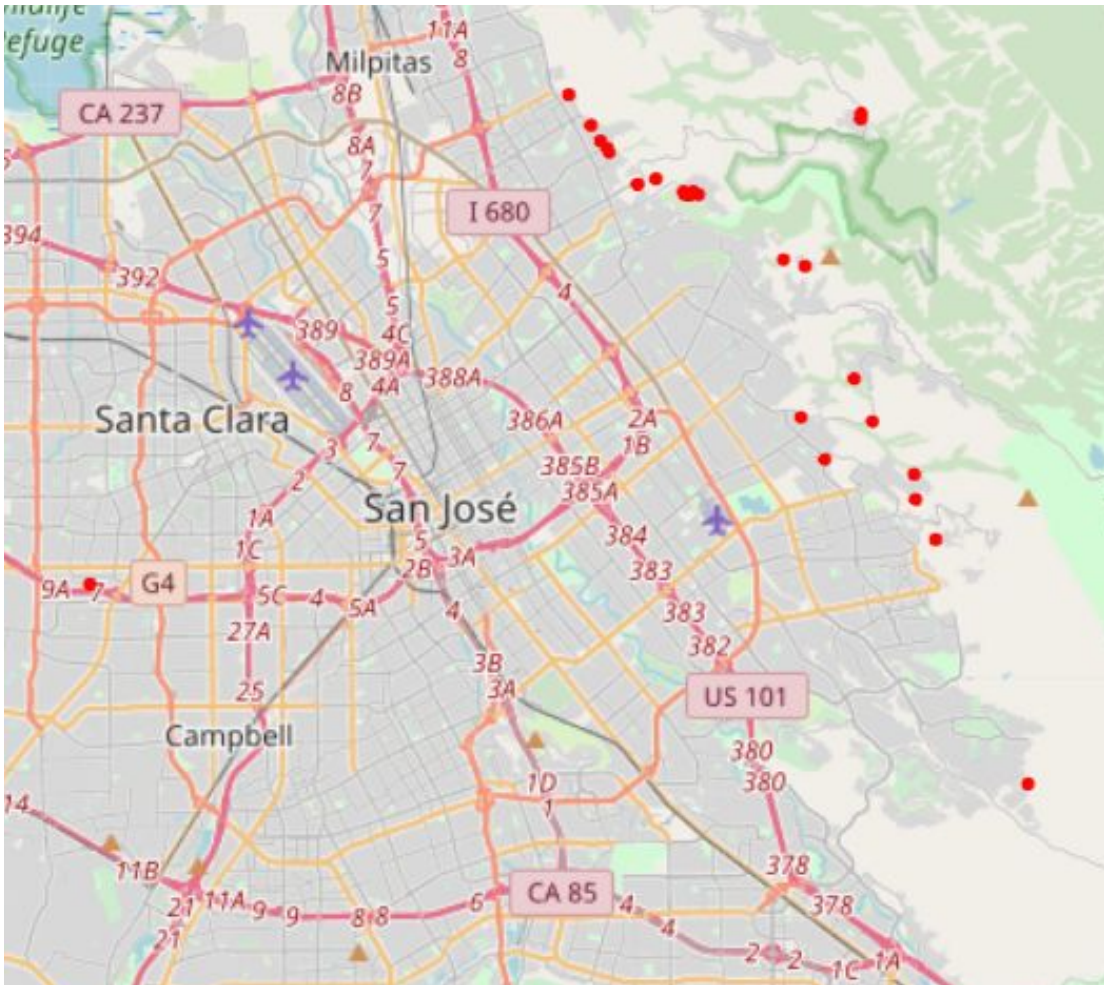


Fig 4: Liquefaction and landslide distribution (single family)

Liquefaction zones are most common hazard in san jose and when it combines with landslide, it is even worse. Among collected data points there were no combined hazard of landslide, liquefaction, landslide. These are points collected and not entire distribution of san jose.

After geospatial analyses, it was decided to do explore number of houses in each category of features.Bar plots were plotted with grouped data to see the distribution of the data.

How many number of bedrooms are most popular?

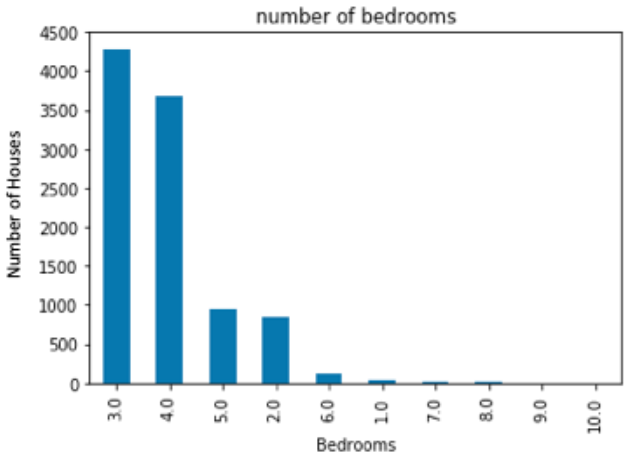


Fig 5a: Bedrooms vs count (single family)

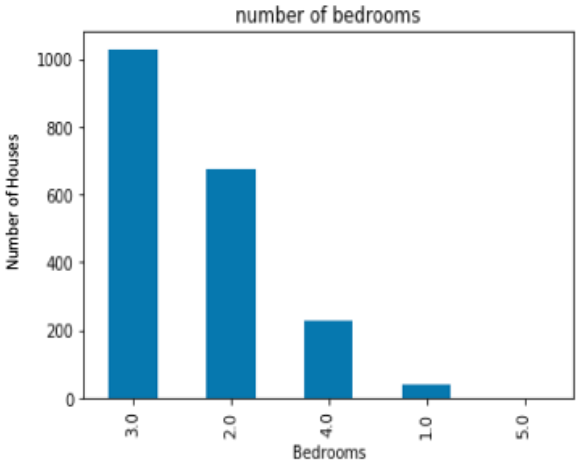


Fig 5b: Bedrooms vs count (townhomes)

The distribution of “number of bedrooms with house count” bar plot shows most popular number of bedrooms is 3 for both single family and townhomes (Figs 5a & 5b).

How many number of bathrooms are most popular?

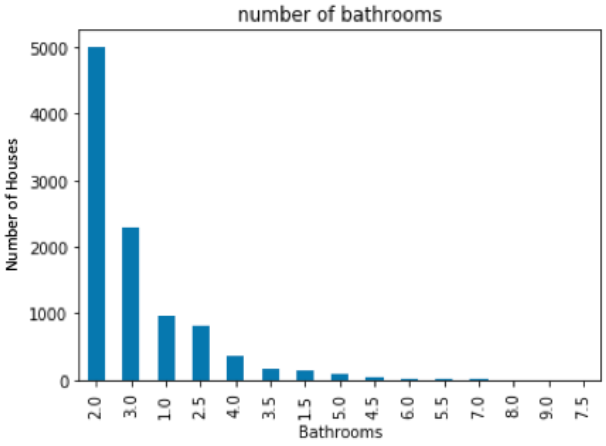


Fig 6a: Bathrooms vs count (single family)

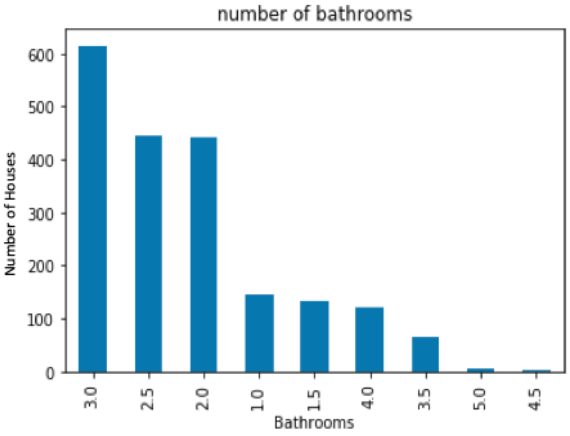


Fig 6b: Bathrooms vs count (townhomes)

The distribution of “number of bathrooms with house count” bar plot shows most popular number of bathrooms is 2 for single family and 3 for townhomes (Figs 6a & 6b)

Which zip code is popular among buyers?

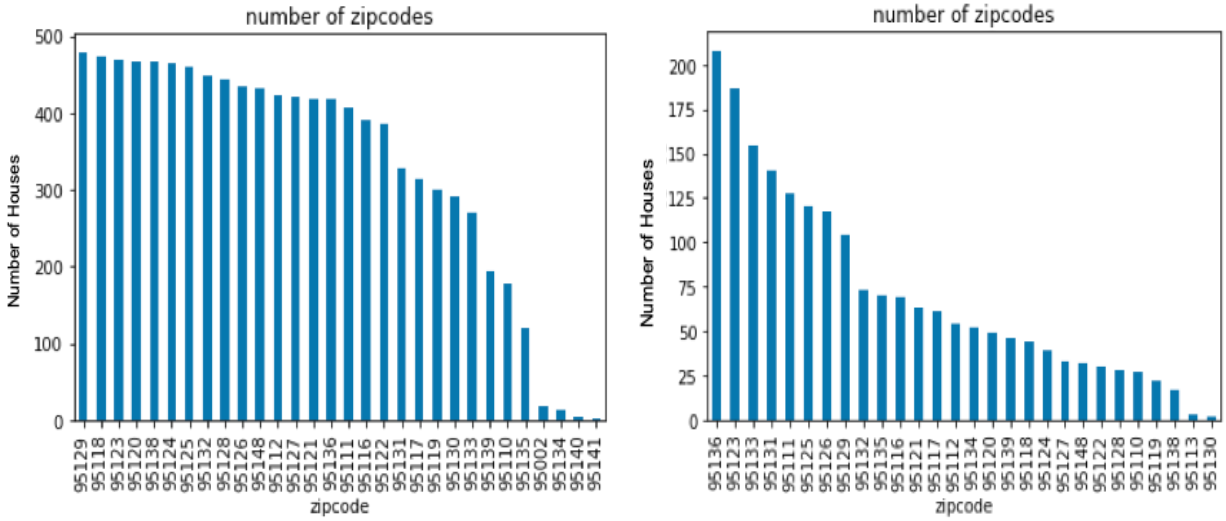


Fig 7a: Zip Code vs Number of Houses (single family) Fig 7b: Zip Code vs Number of Houses (townhomes)

Among 30 zip codes, there are more than dozen zip codes are equally popular for single family homes. For townhomes, 8 zip codes are more popular (Figs 7a & 7b)

Which year built is the most common ?

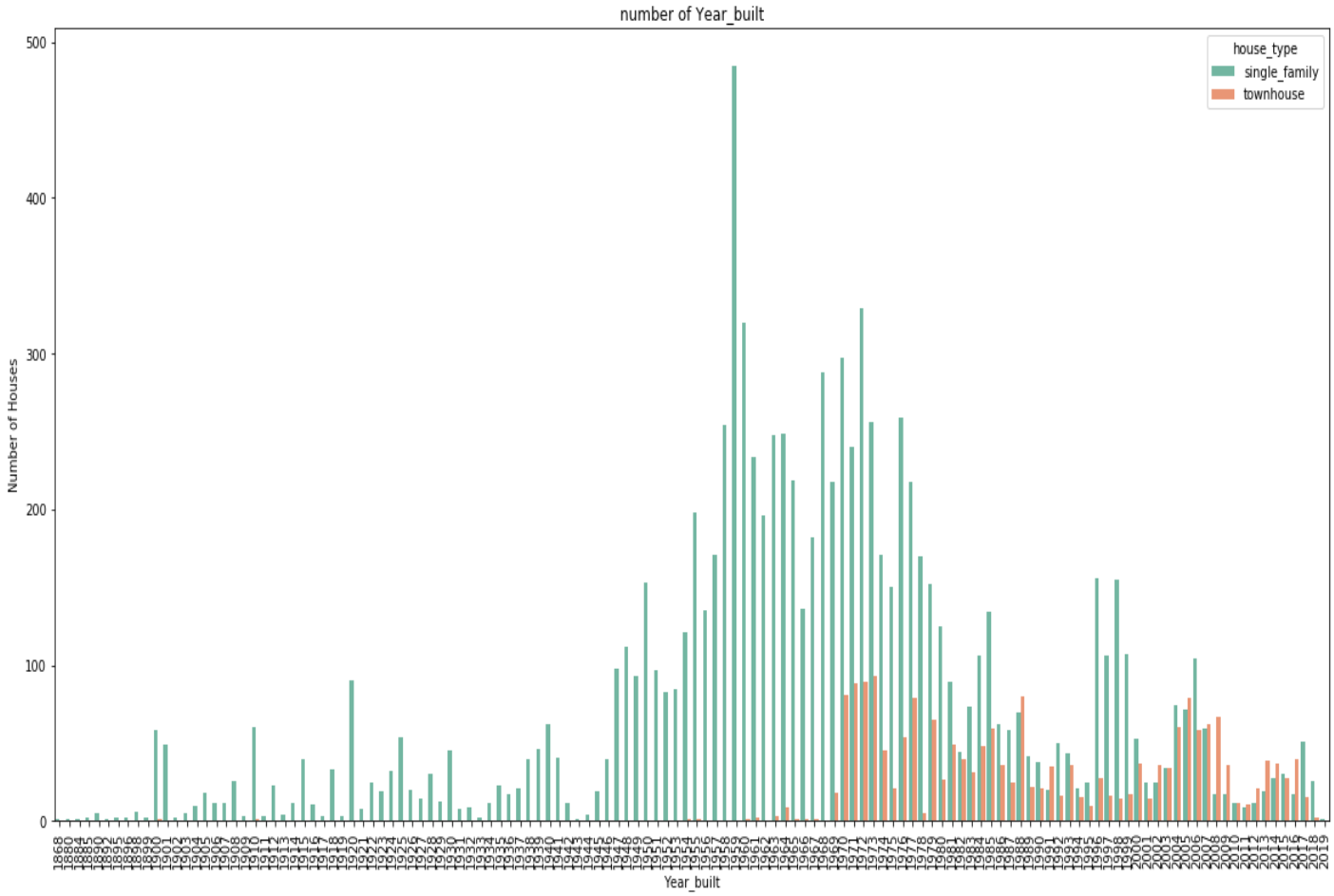


Fig 8: Year built vs Number of Houses (Single and townhomes)

The distribution of “year built with house count” bar plot shows most common year built is 1959 for single family. For townhomes, most of them were built during 1988-1973 and 2005 & 2008 (Figs 8).

Which month is popular for selling house?

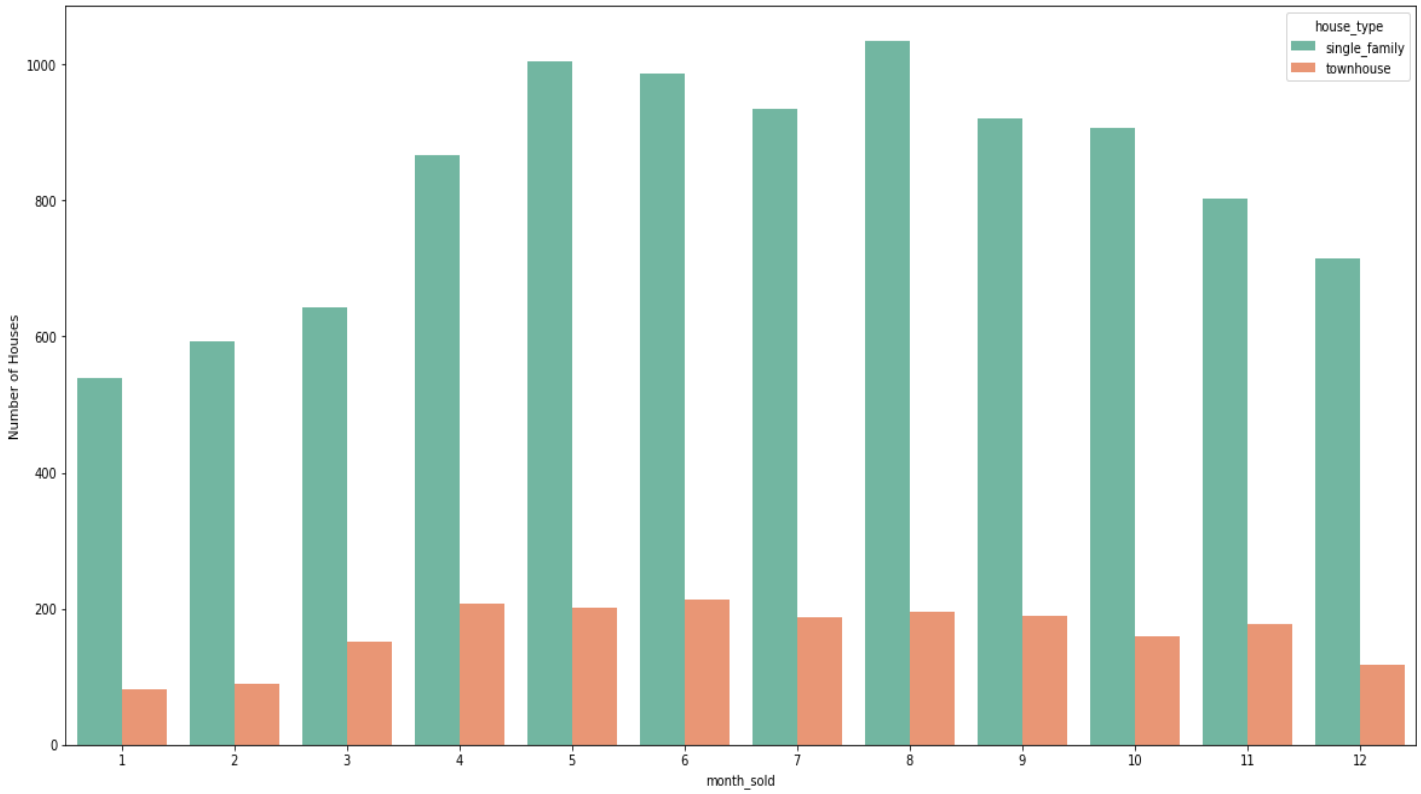


Fig 9: Month_sold vs Number of Houses (single family & townhouse)

The distribution of “month_sold with house count” bar plot shows best month to sell house is during May-August for both single family and townhomes. This is because, during summer break, most people plan to relocate (Fig 9).

Which year was most house solded?

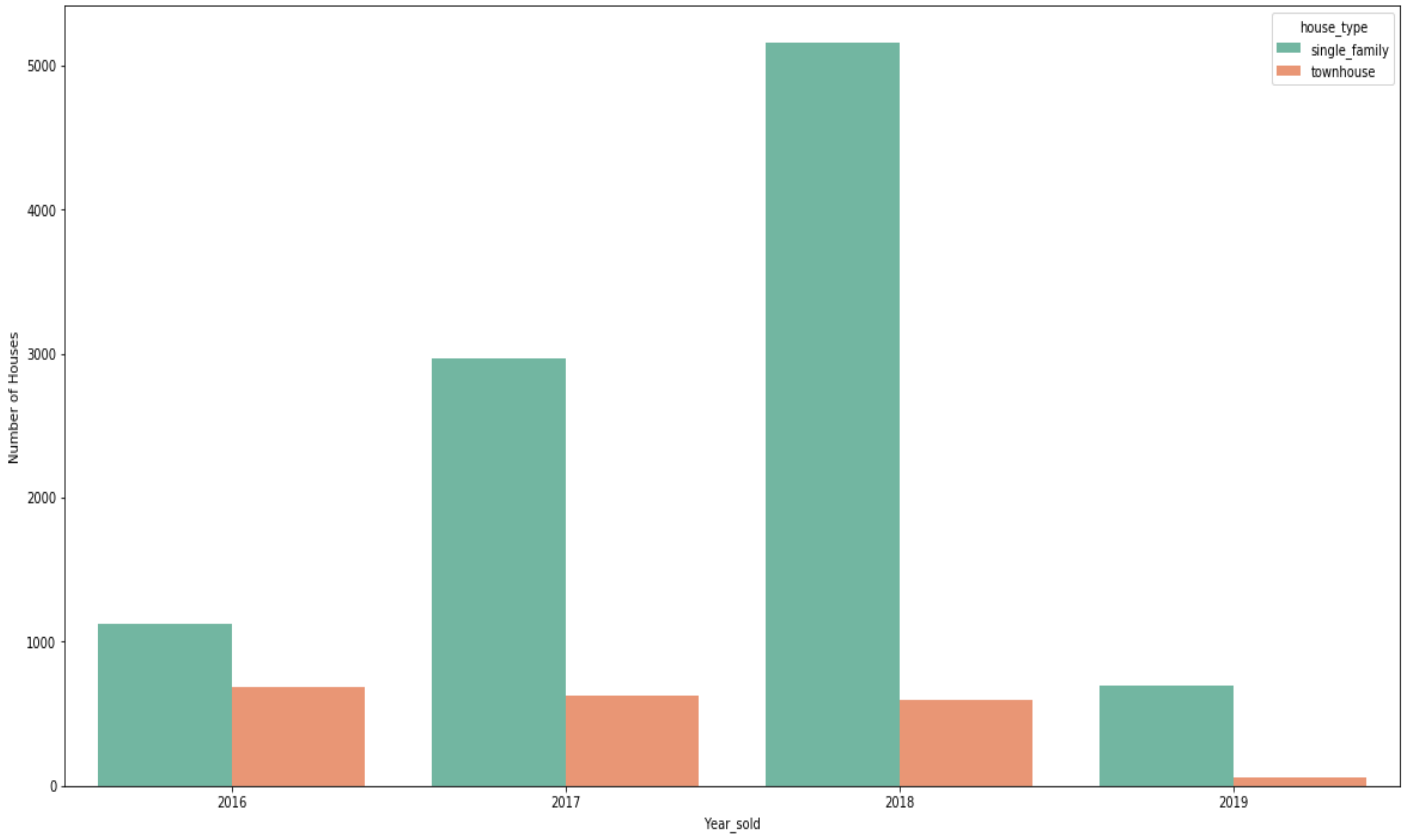


Fig 10: Year_sold vs Number of Houses(single family & Townhouse)

The distribution of “year_sold with house count” bar plot shows that most of the houses were sold in 2018 for single family and 2016-2018 for townhomes (Fig 10).

Which hazard is most common?

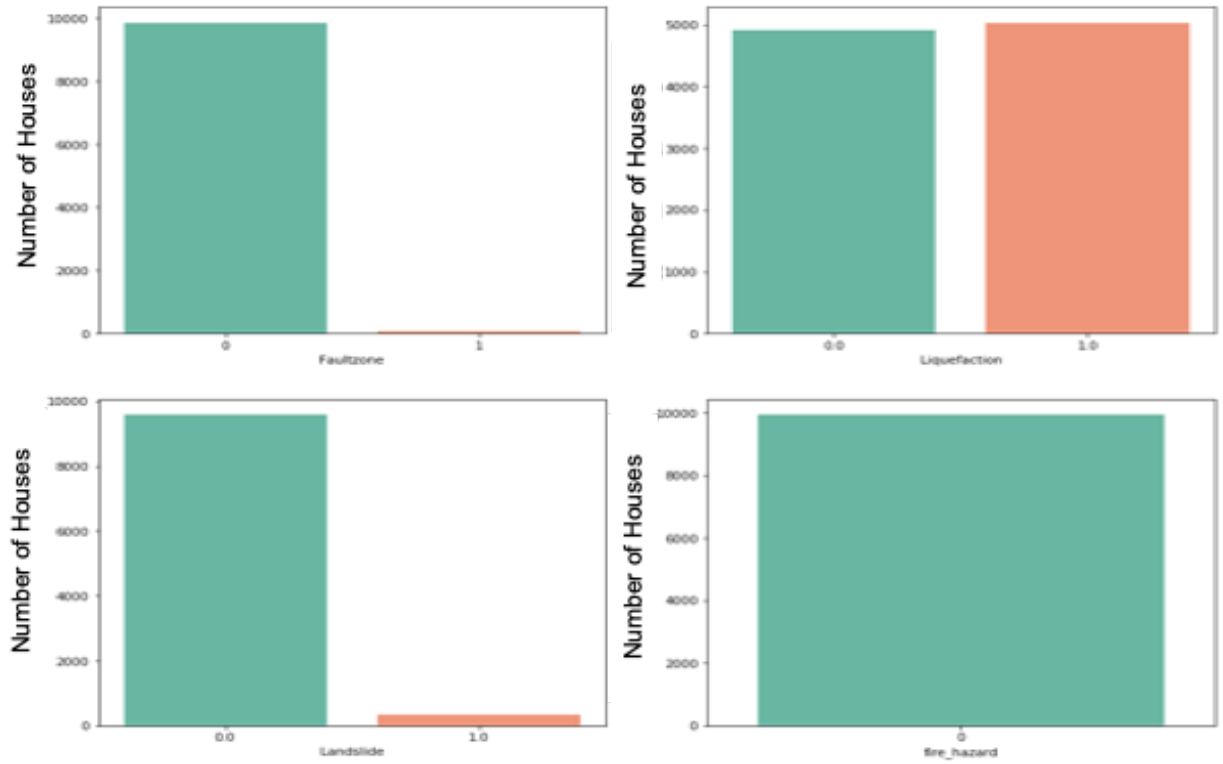


Fig 11a: Hazard vs Number of Houses (single family)

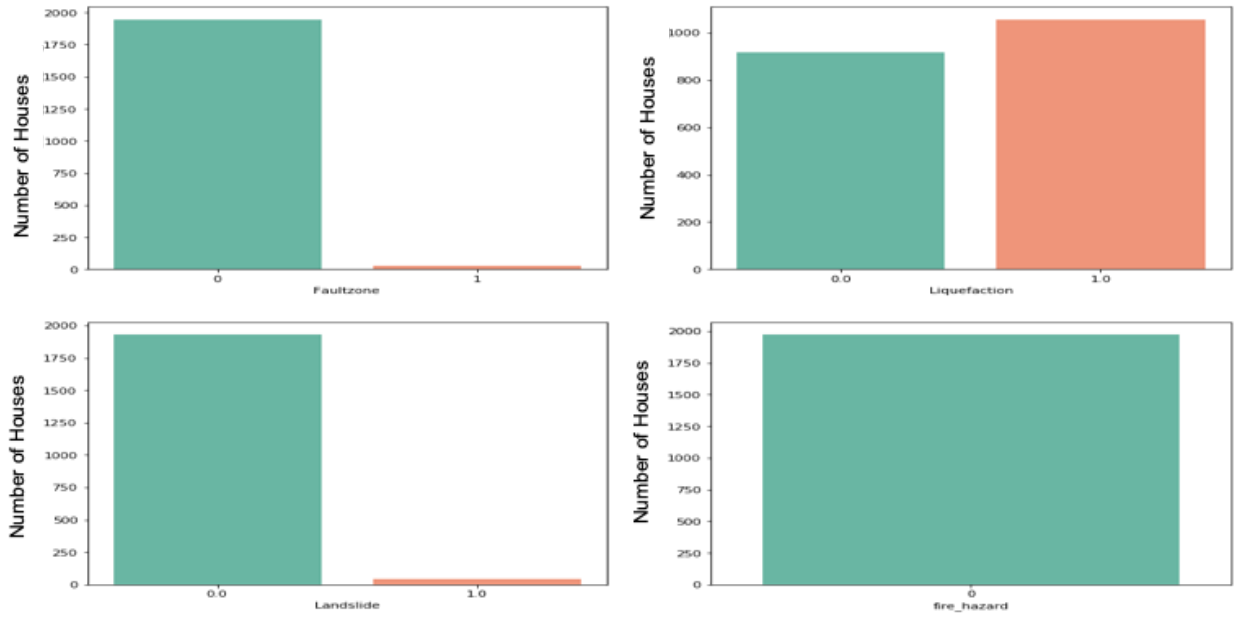


Fig 11b: Hazard vs Number of Houses (townhomes)

The distribution of “hazard with house count” bar plot shows that most of the houses were vulnerable to liquefaction for both single family and townhomes (Figs 11a & 11b).

How many number of houses in different price bin?

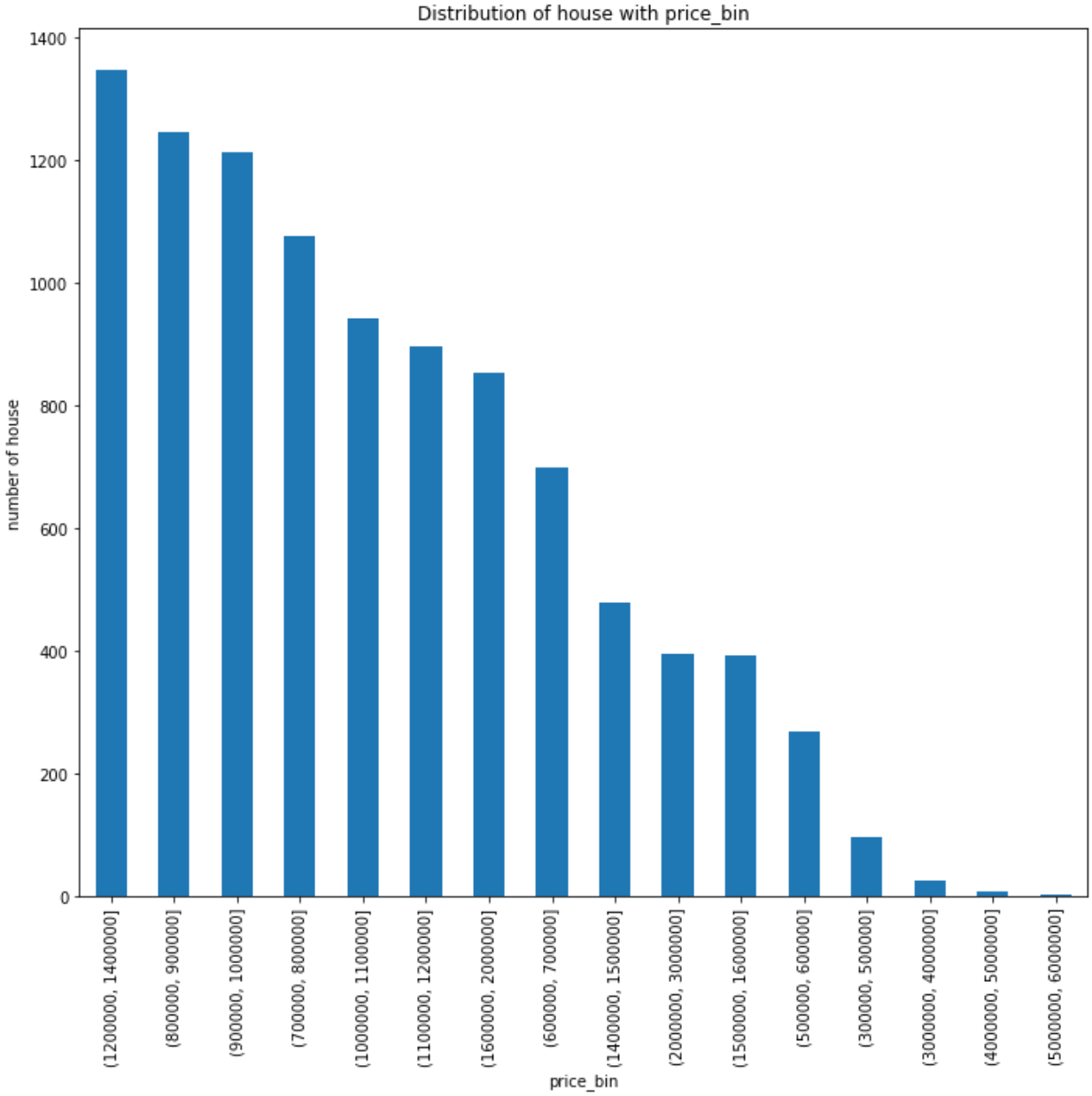


Fig 12a: Price_bin vs count (single family)

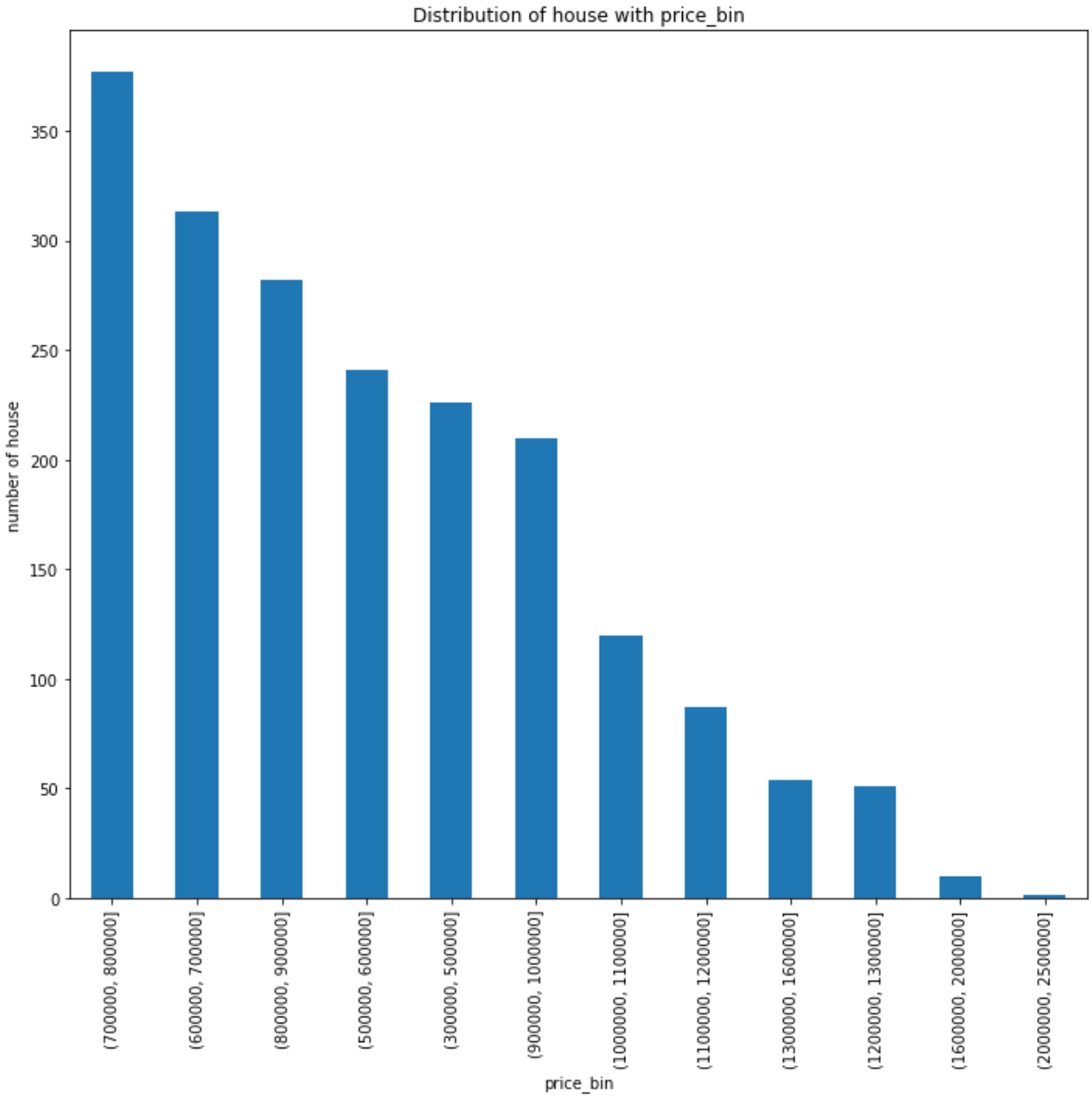


Fig 12b: Price_bin vs count (townhomes)

How many number of bedrooms in different price bin?

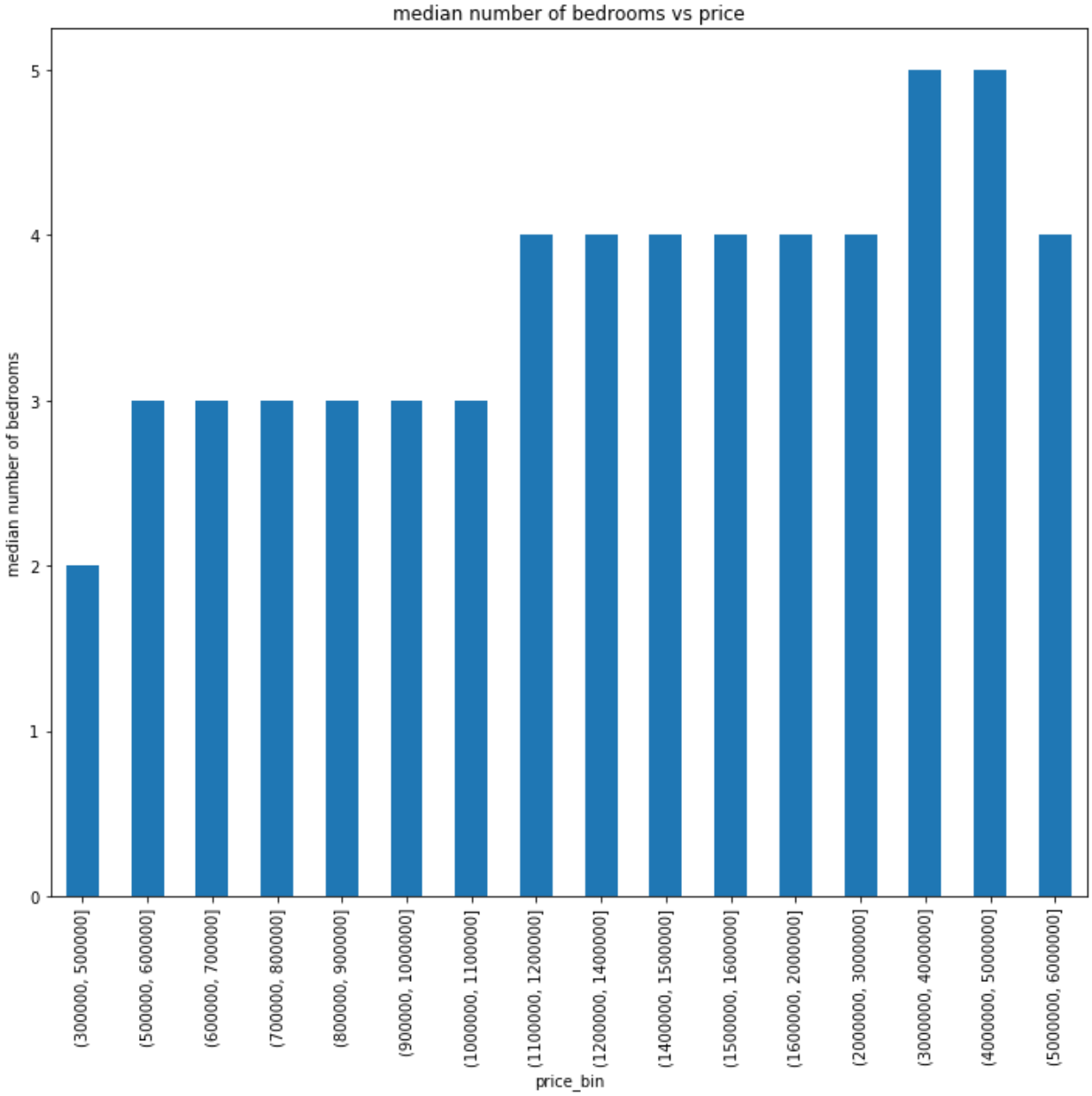


Fig 13a: Price_bin vs median bedrooms (single family)

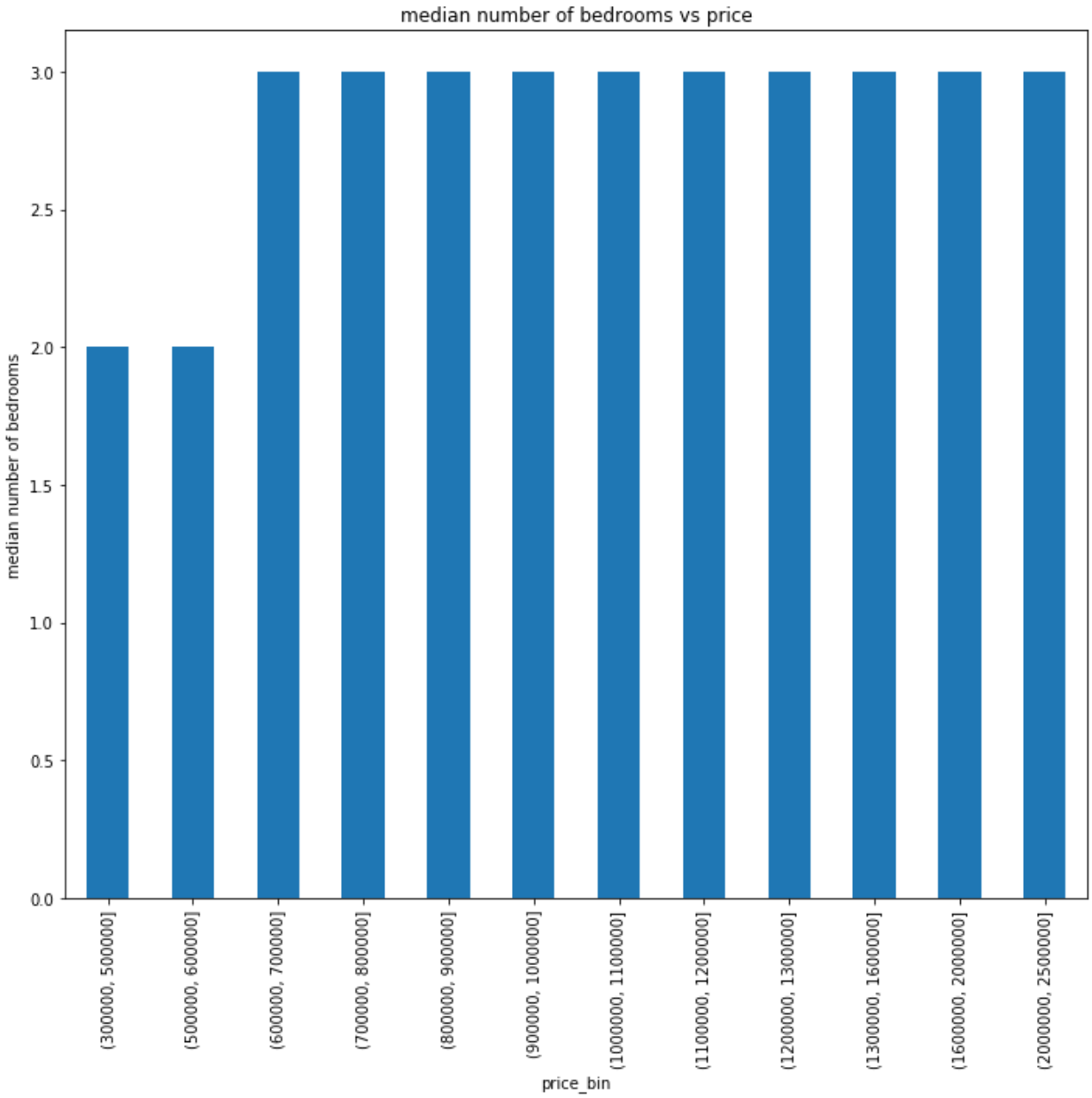


Fig 13b: Price_bin vs median bedrooms (townhomes)

How many number of bathrooms in various price bin?

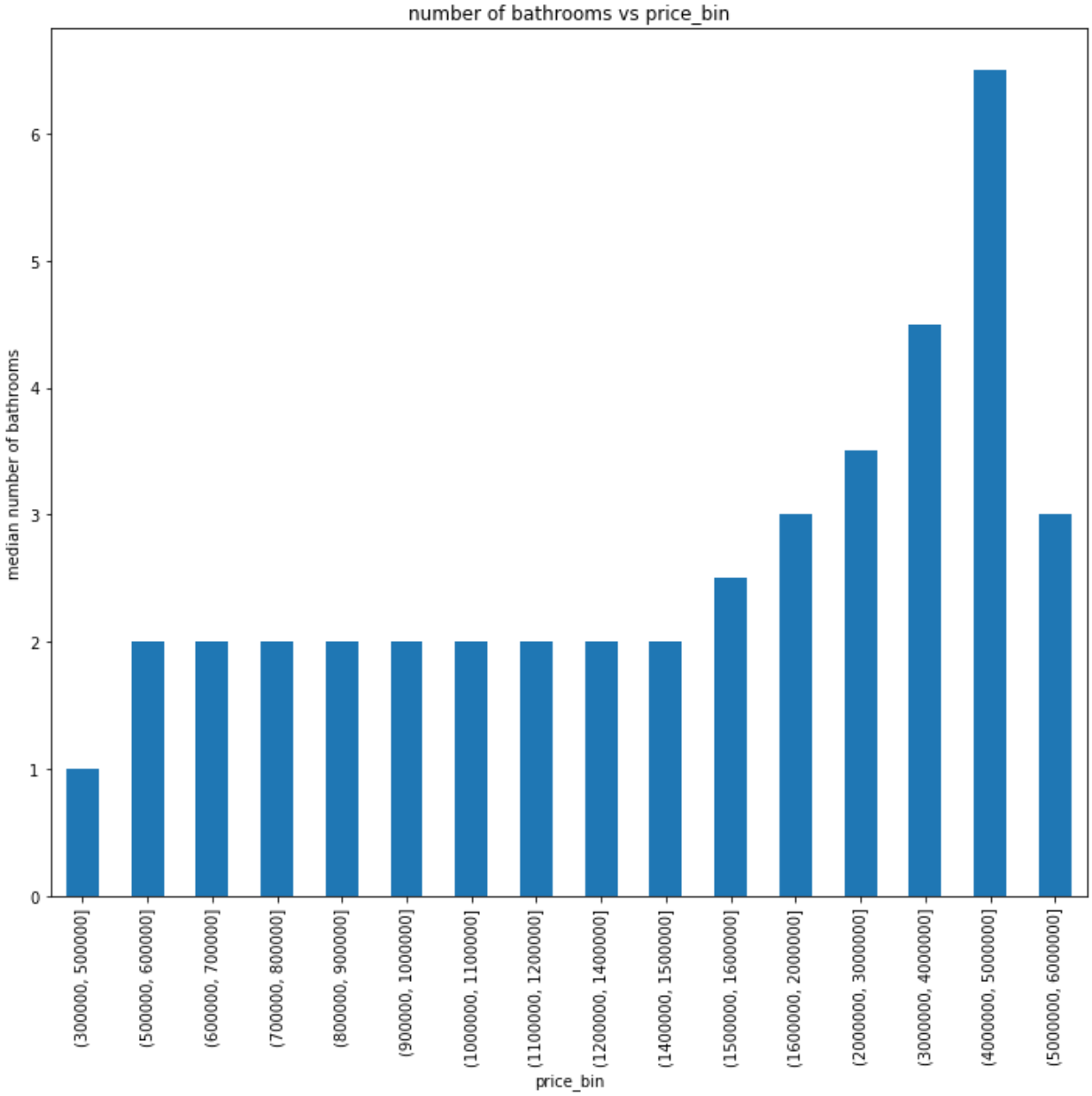


Fig 14a: Price_bin vs median bathrooms (single family)

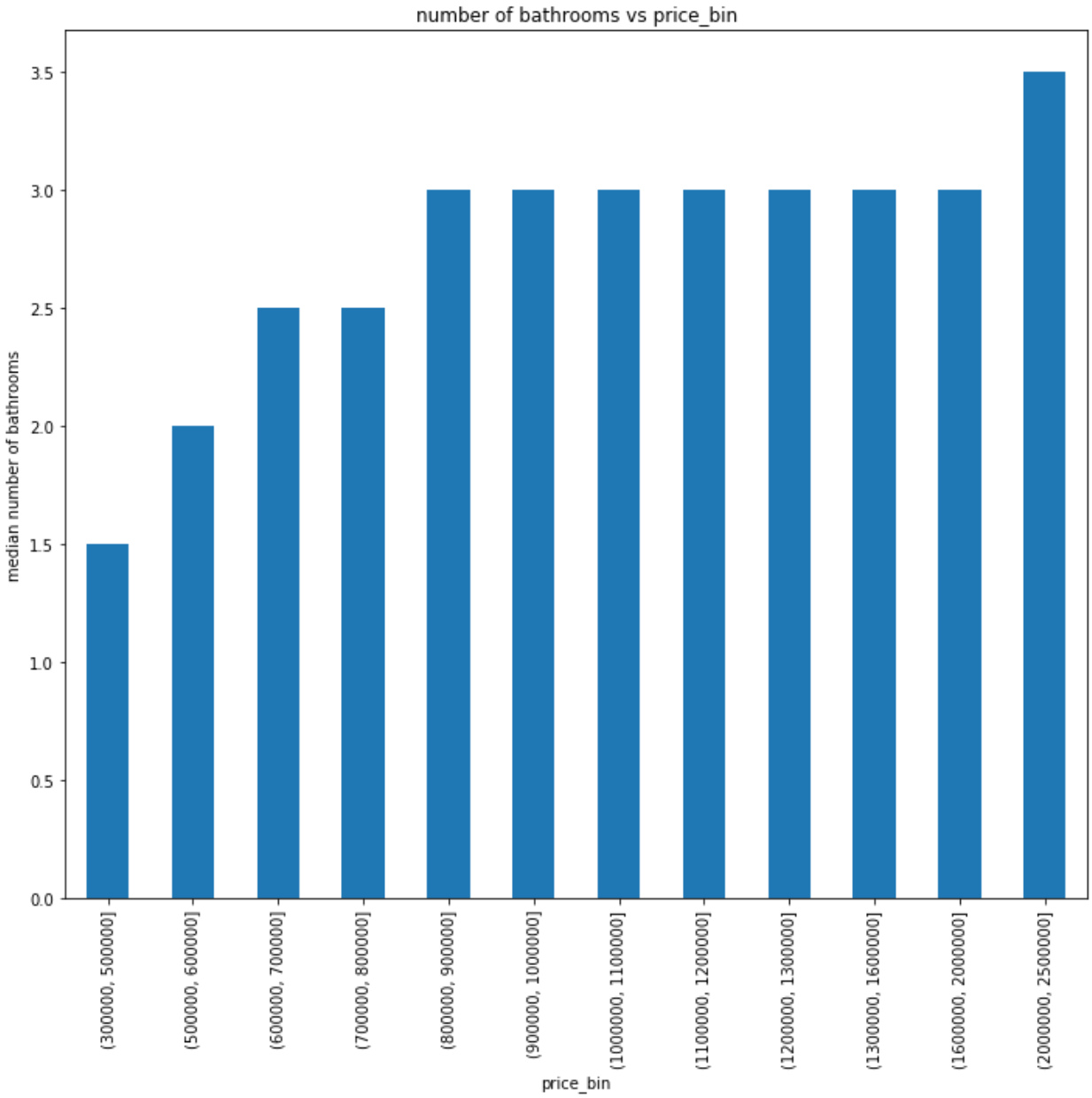


Fig 14b: Price_bin vs median bathrooms (townhomes)

Which median sqft is common within various price bin?

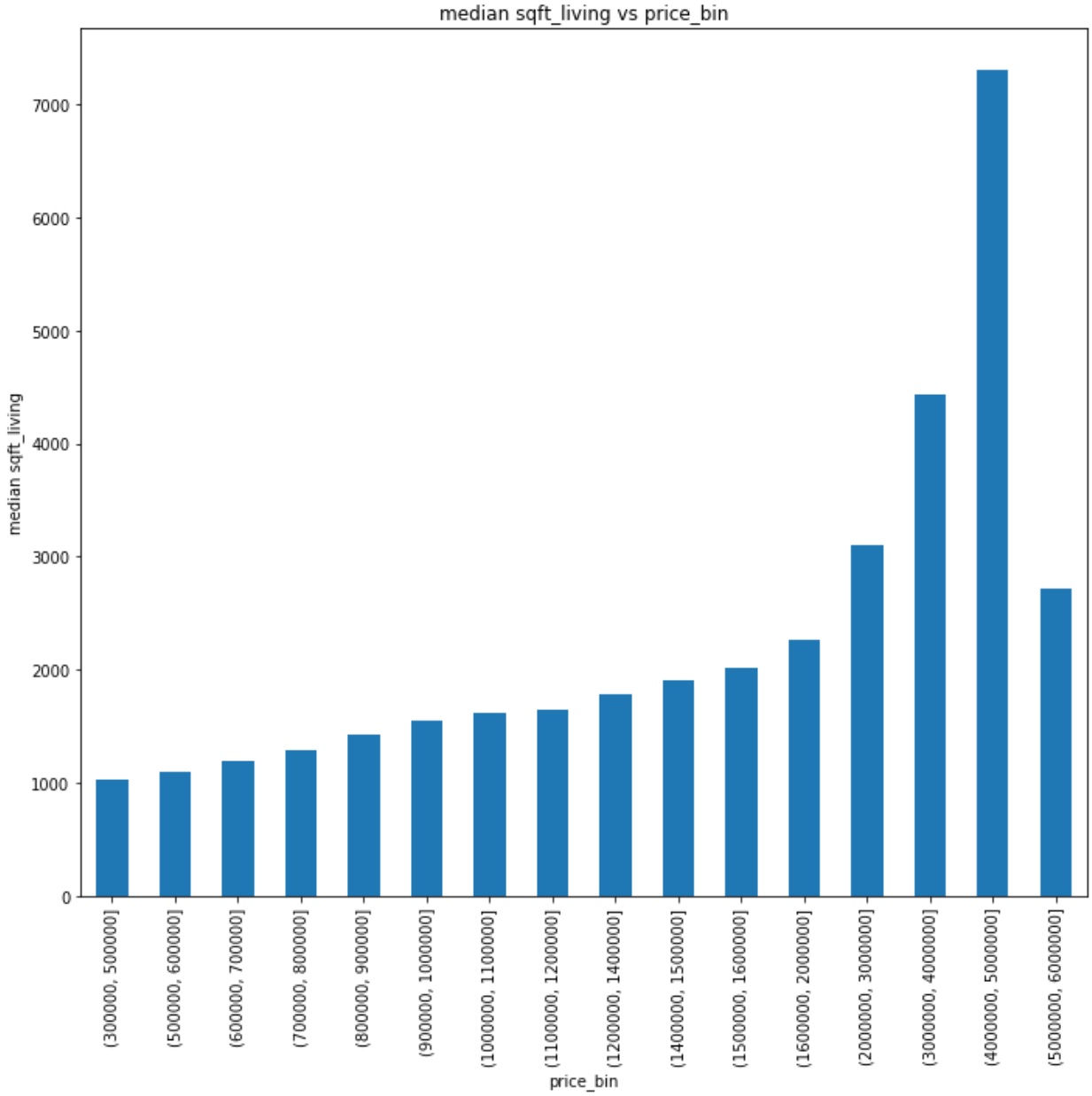


Fig 15a: Price_bin vs median sqft (single family)

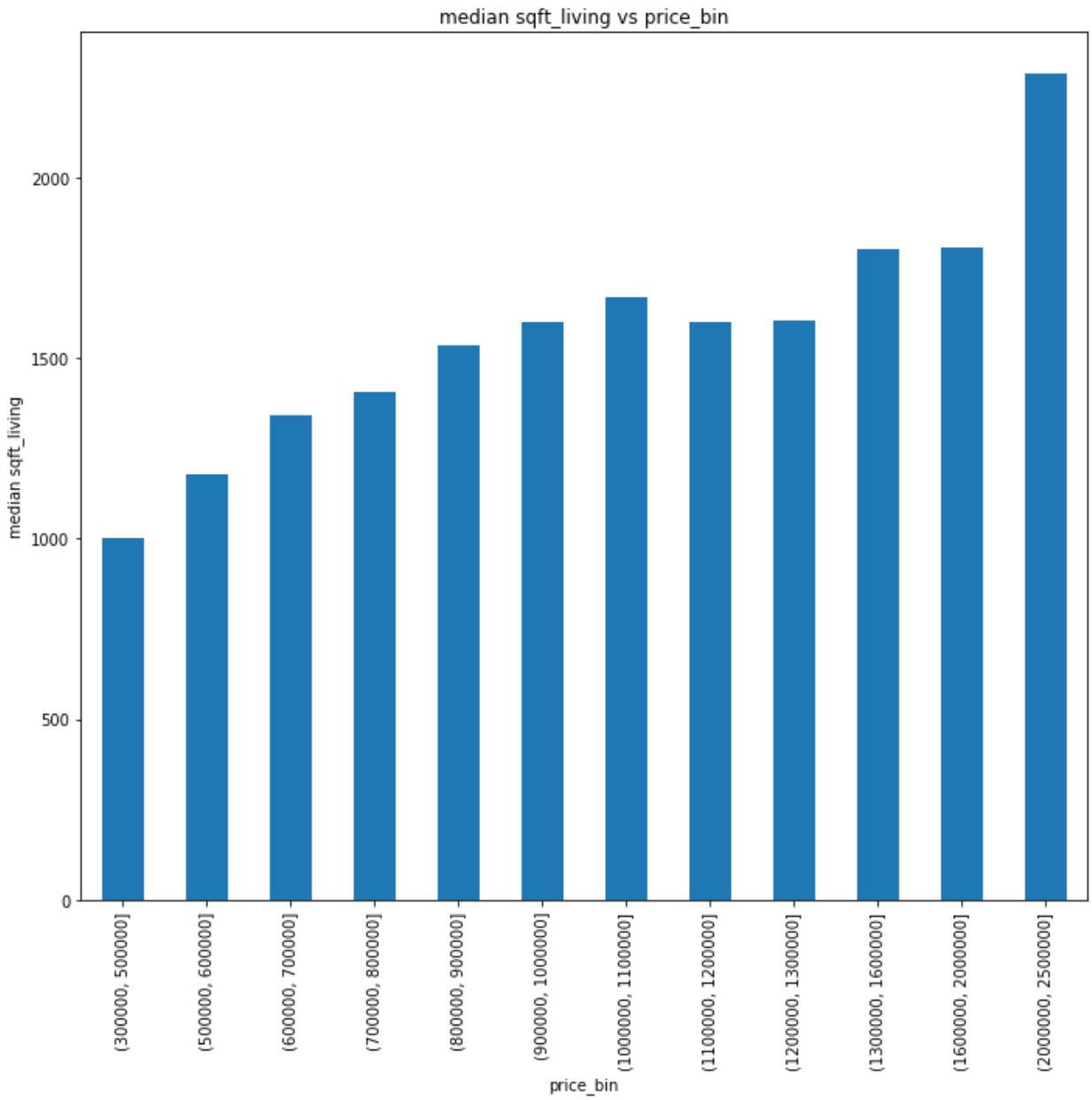


Fig 15b: Price_bin vs median sqft (townhomes)

Which price range homes are in liquefaction zone?

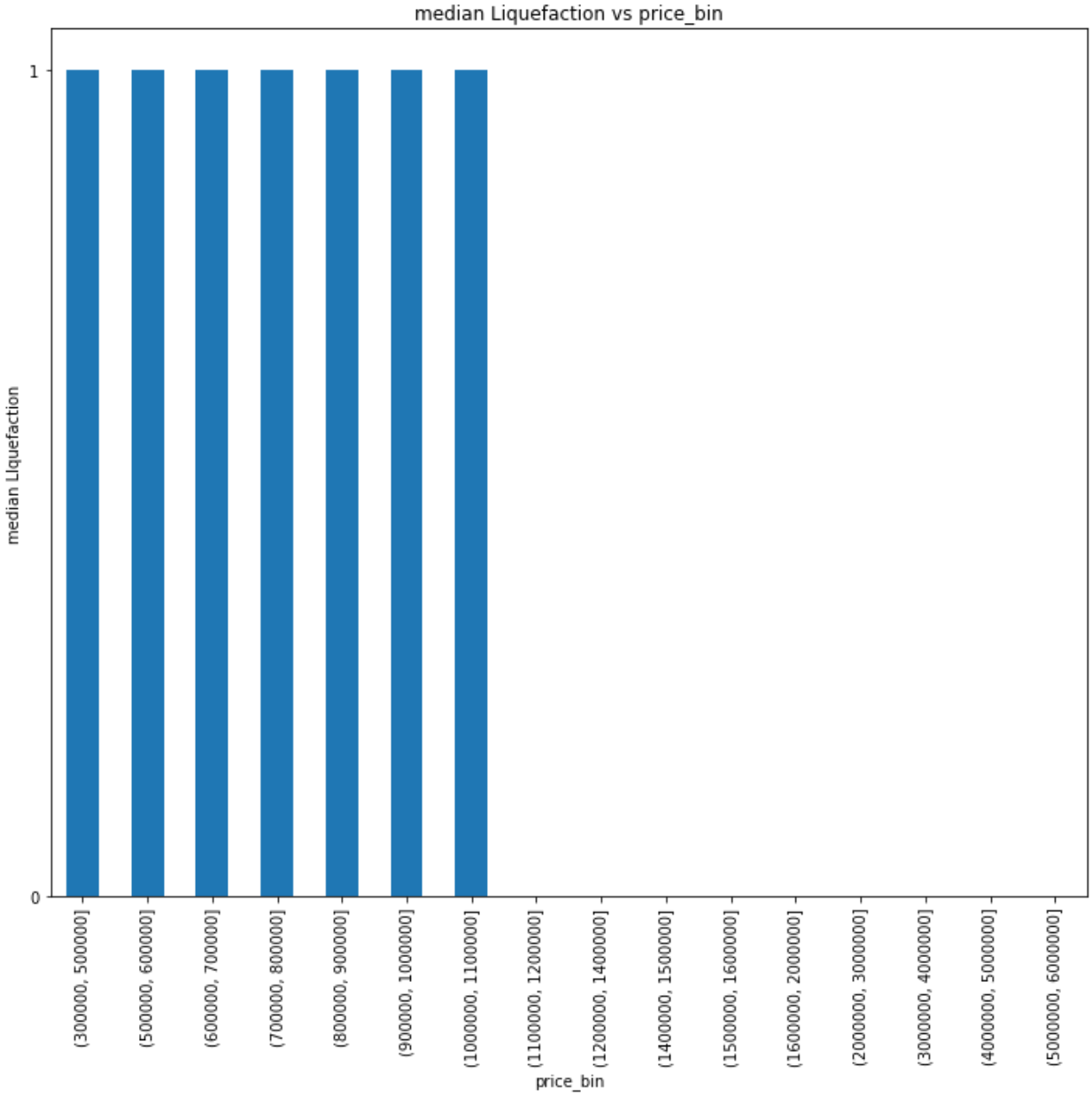


Fig 16a: Price_bin vs liquefaction (single family)

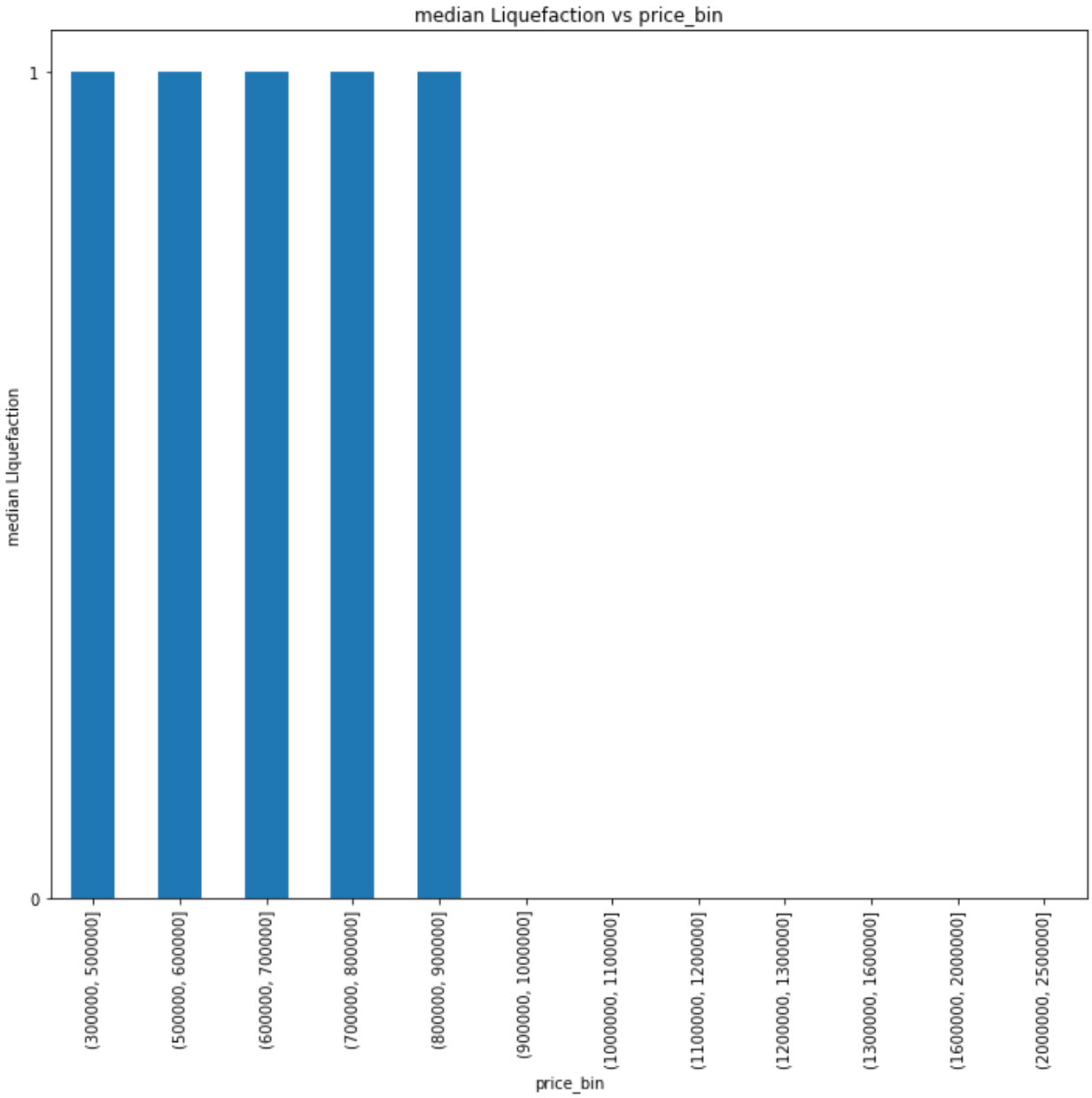


Fig 16b: Price_bin vs liquefaction (townhomes)

Which price range homes are in landslide zone?

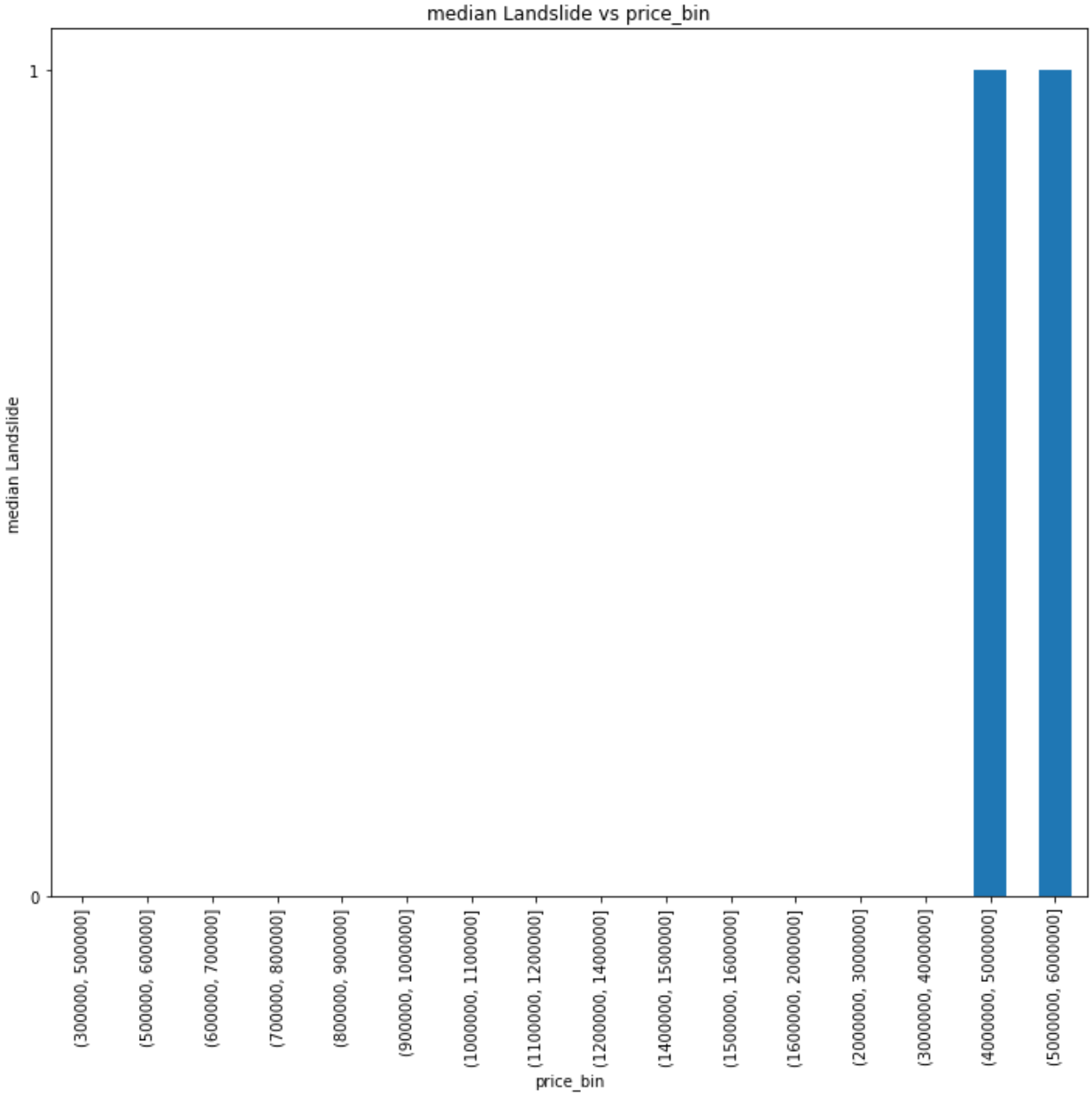


Fig 17: Price_bin vs landslide (single family)

Price distribution for various zip codes, hazard and non hazard zones

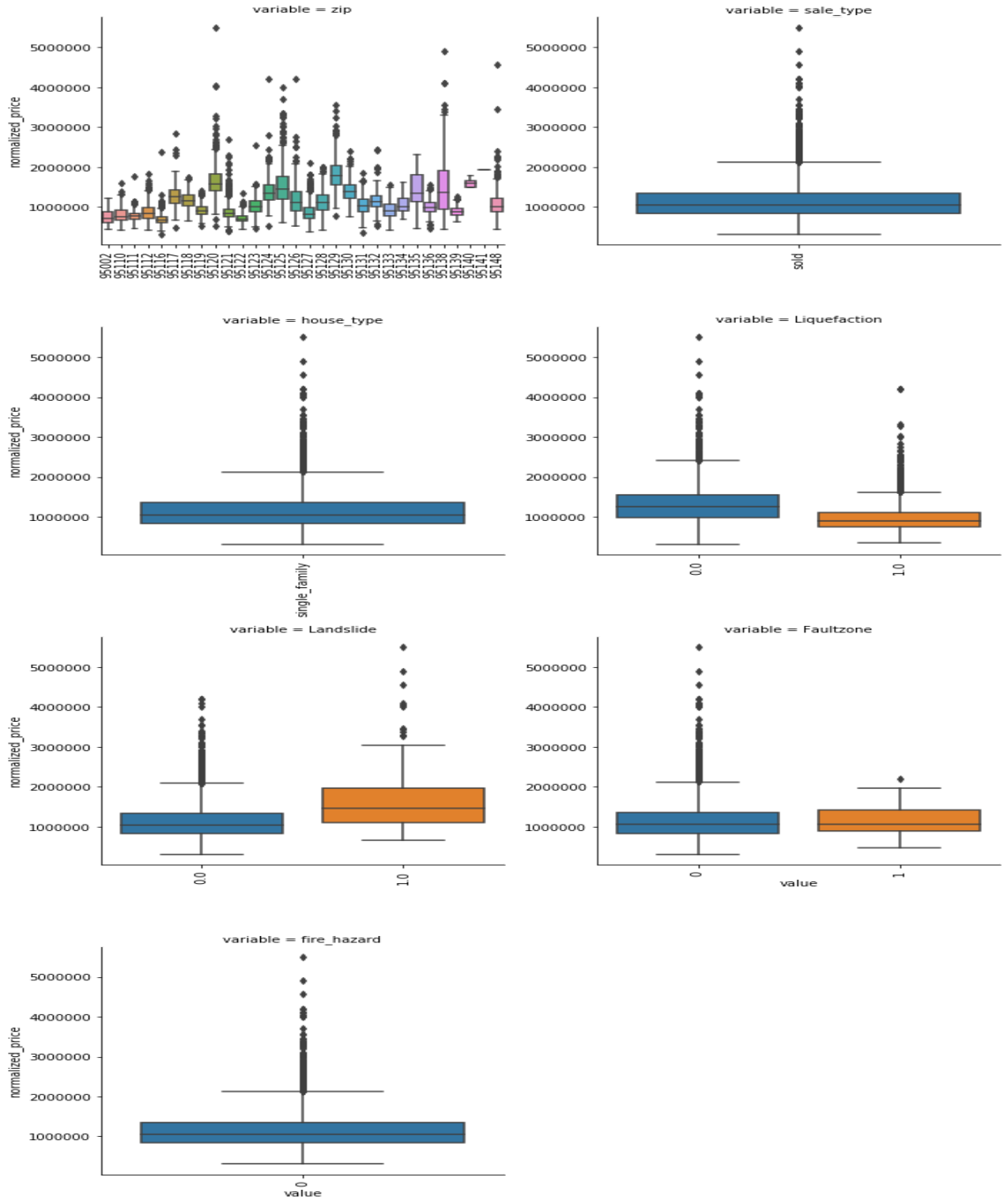


Fig 18a: Price distribution box plot (single family)

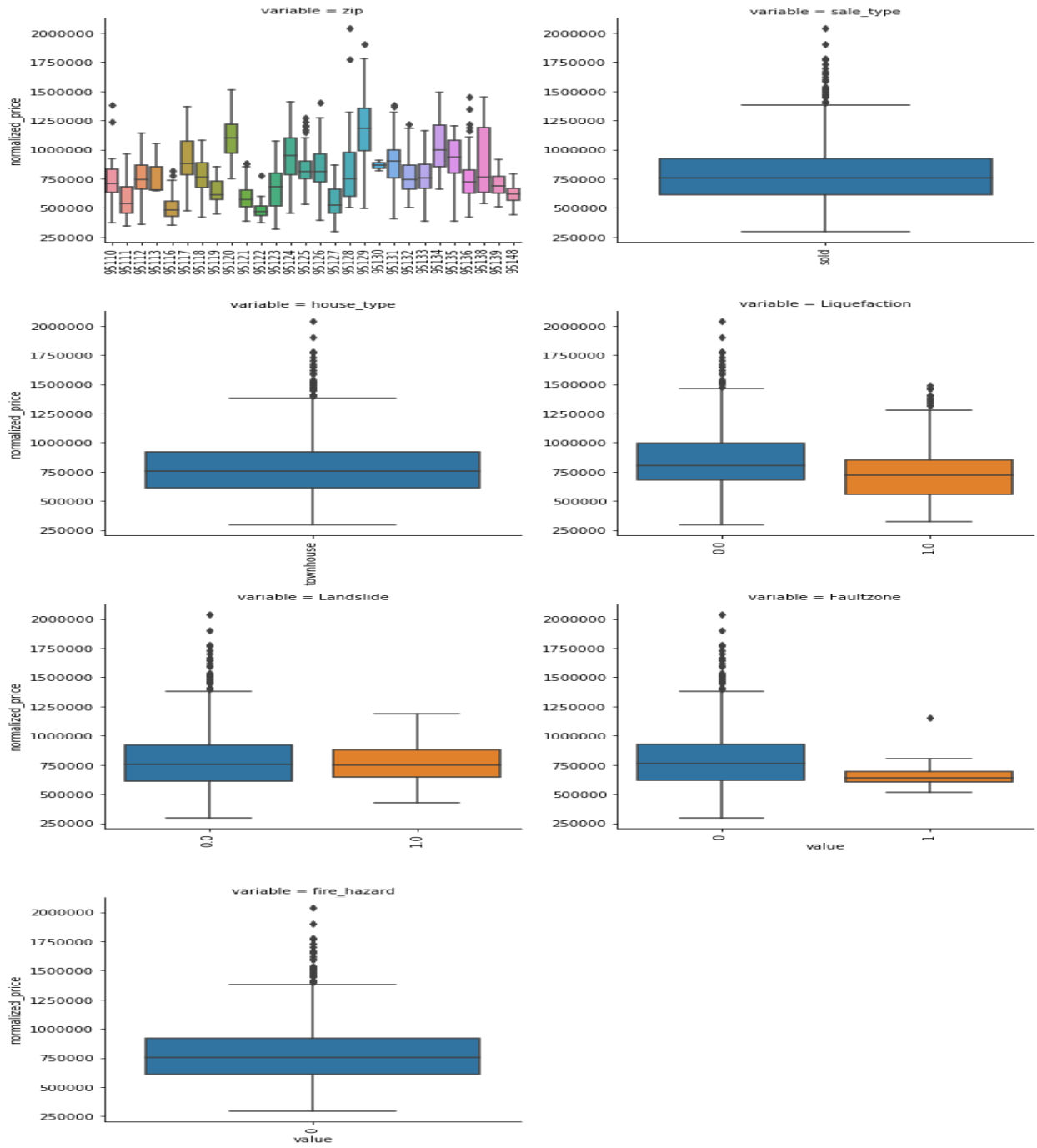


Fig 18b: Price distribution box plot (townhomes)

Above plot clearly shows the effect of natural hazards on the price of the house. For example, the liquefaction prone areas are less price compared to non liquefaction areas (in 4th plot). Similarly, fault zone areas are less price than no fault zone areas (in 6th plot). But it is contradicting for landslide areas (in 5th plot). Landslide areas are predominantly in mountain areas. This indicates people are preferring spectacular views rather than landslide hazard when buying a house. More details follow in subsequent sections.

What is the effect of natural hazard (Liquefaction) in price?

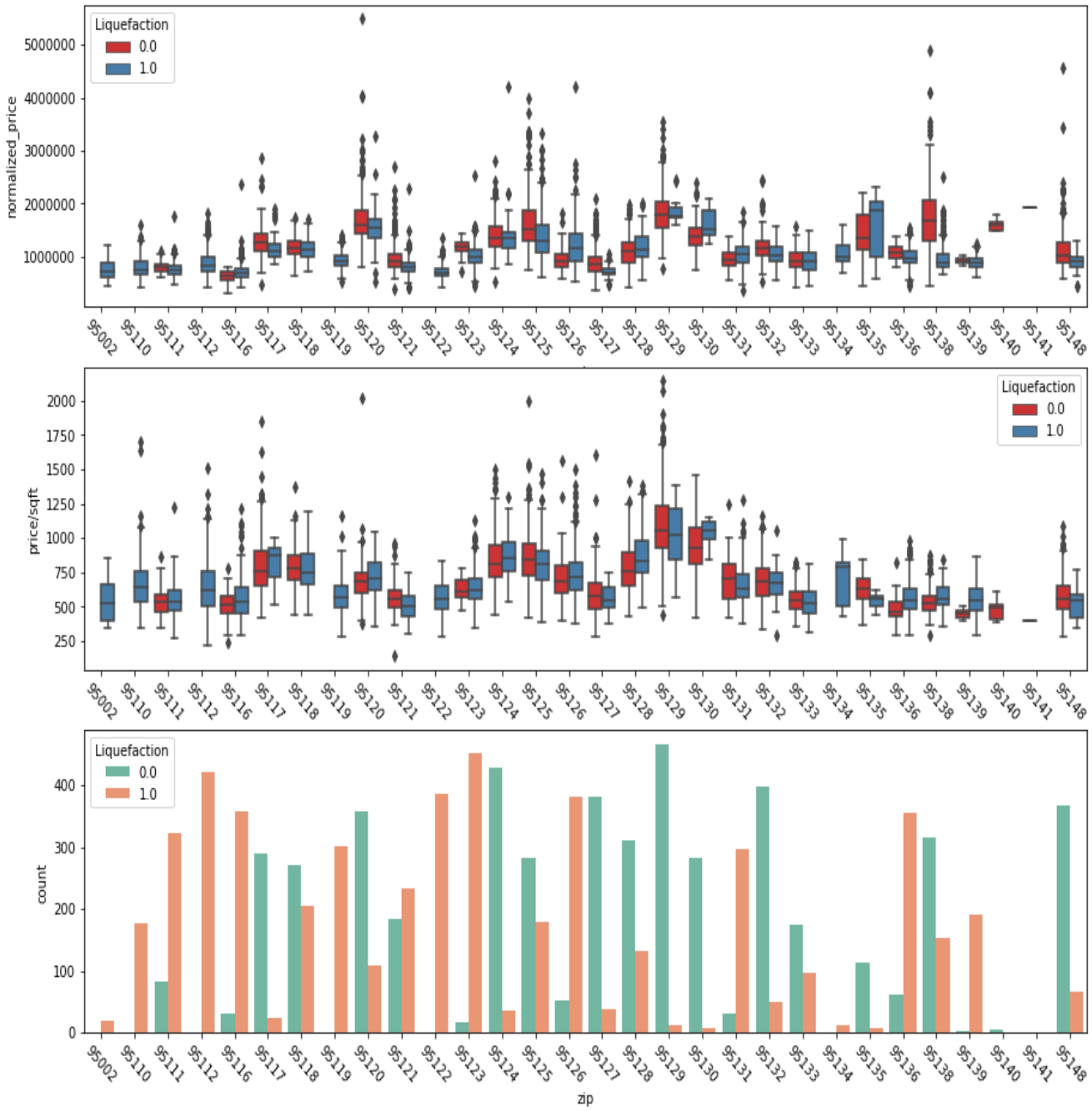


Fig 19a: Price distribution box plot (single family)

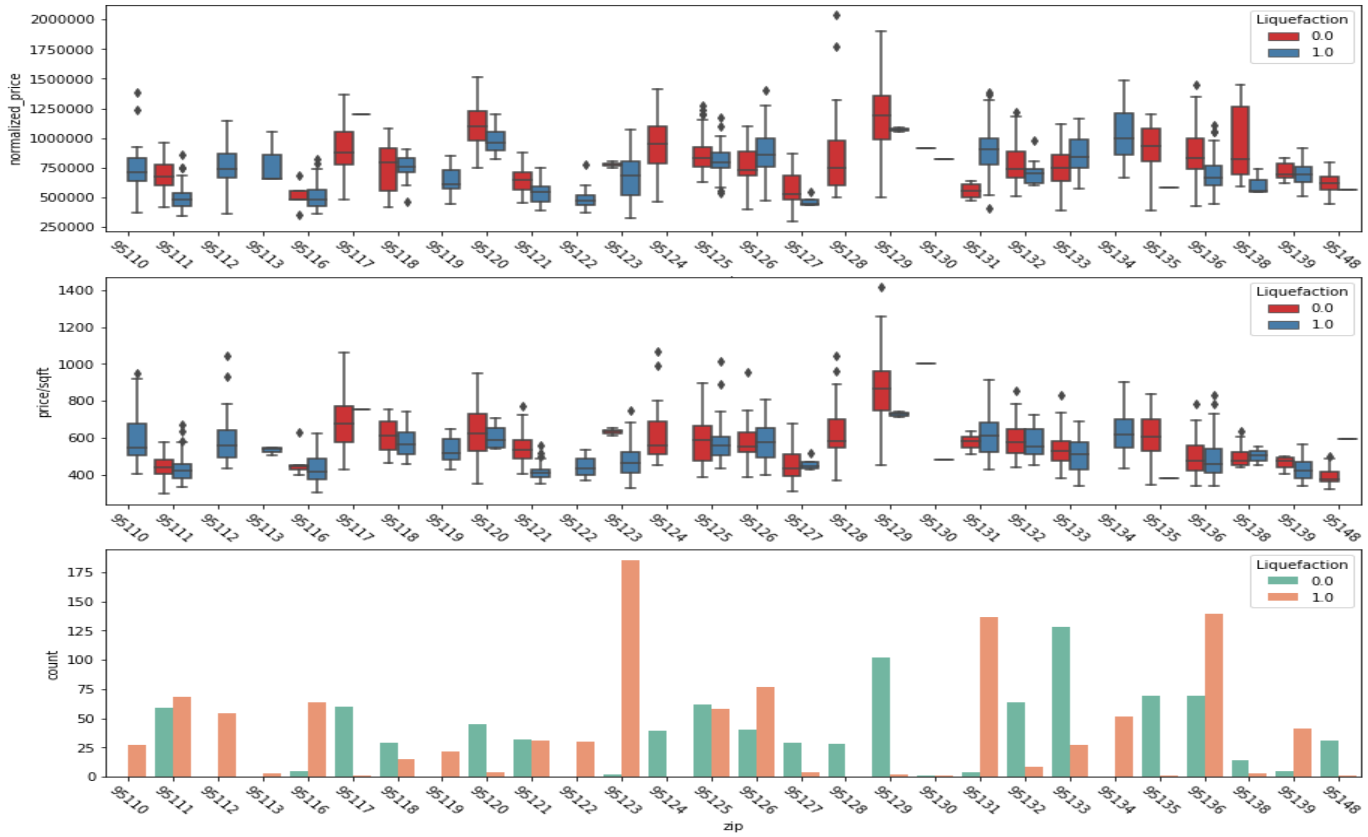


Fig 19b: Price distribution box plot (townhomes)

Above plot shows price distribution, price/sqft, count for hazard and non hazard various zip codes in single family homes. In order to study the effect of liquefaction on price, equal number of hazard and non hazard homes were analyzed in nearby zip codes with equal price /sqft. For this, zip code 95116 was considered, with liquefaction hazard and 95127 with no liquefaction hazard. Below median value table indicates the liquefaction prone homes are selling less than non liquefaction areas. The name of the neighborhood for each zip code are presented in Table [1].

	With liquefaction 95116	Without liquefaction 95127
count	359	382
mean	$708797	$895599
min	$404000	$375500
25%	$615000	$730000
50%	$680000	$845530
75%	$763500	$1.0M
max	$2.38M	$2.1M

What is the effect of natural hazard (Landslide) in price?

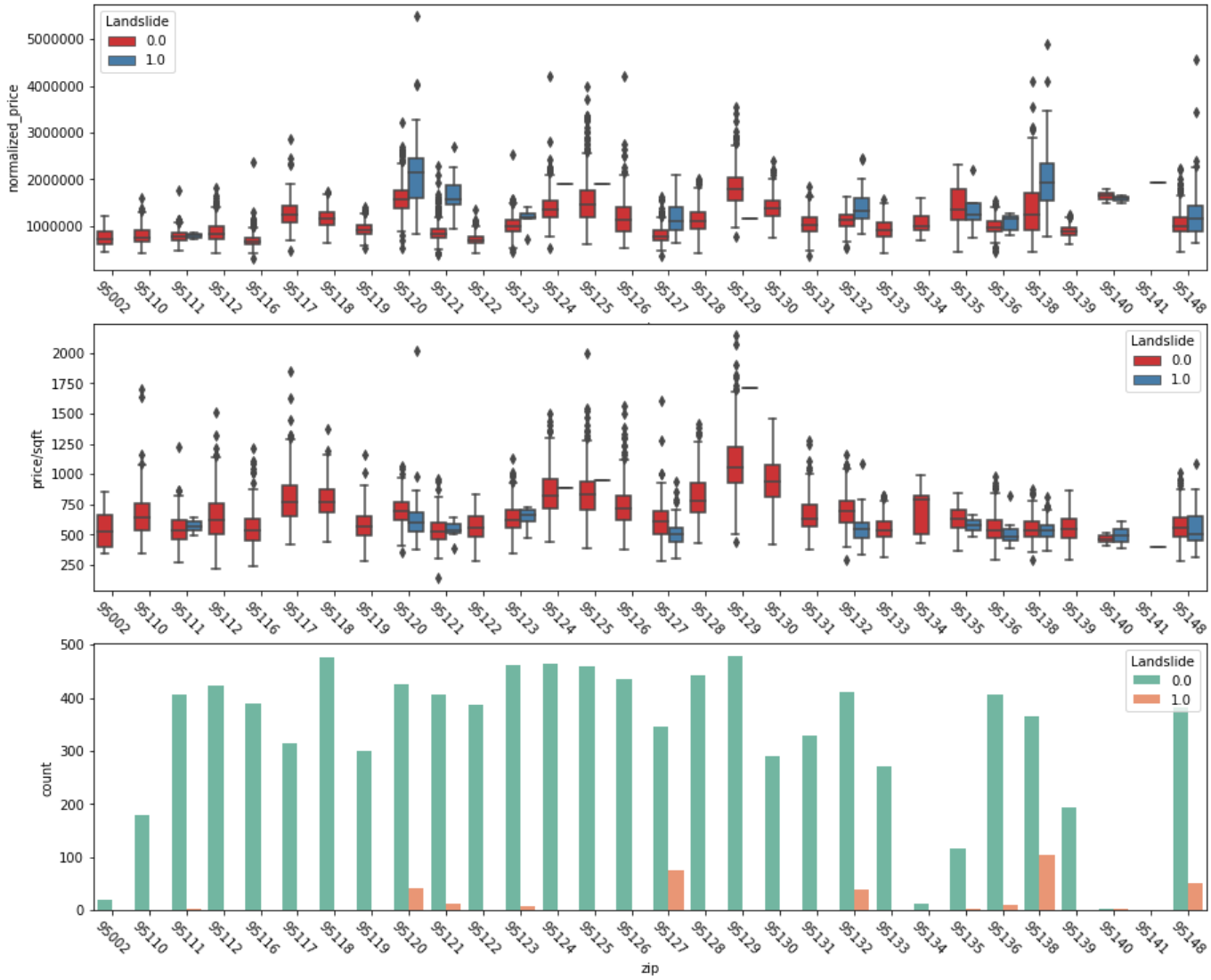


Fig 20a: Price distribution box plot (single family)

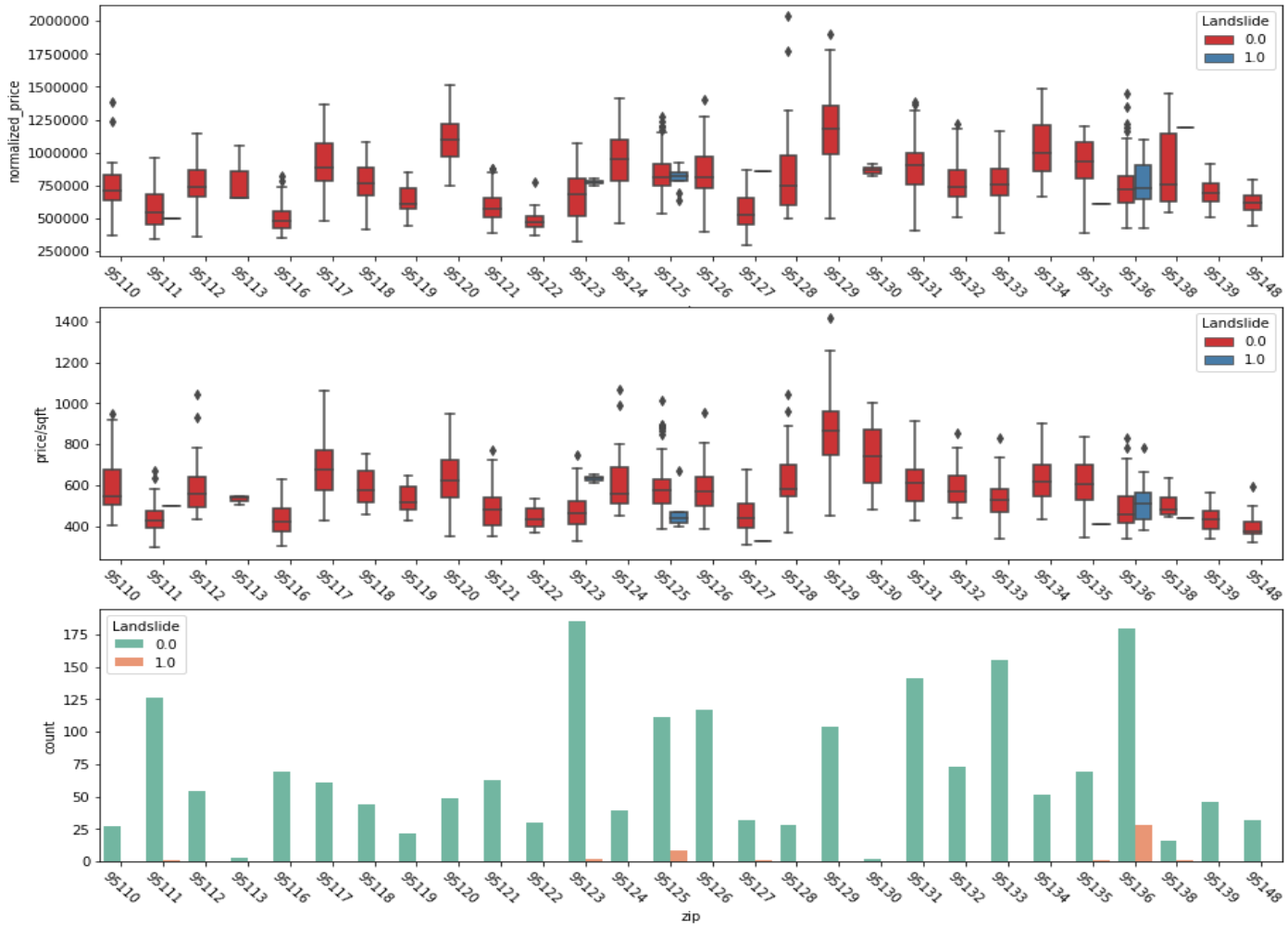


Fig 20b: Price distribution box plot (townhomes)

Above plot shows price distribution, price/sqft, count for hazard and non hazard various zip codes in single family homes. In order to study the effect of landslide on price, equal number of hazard and non hazard homes were analyzed in nearby zip codes with equal price /sqft. For this, zip code 95138 with landslide hazard and 95135 with no liquefaction hazard were considered. Below median price value table indicates the landslide prone homes are selling higher than non landslide areas. This indicates people are preferring spectacular views rather than landslide hazard when buying a house. Sample review comments from 95138 neighborhood to show people's thoughts on this neighborhood were included here.

"Safe and friendly. Near open space including parks, foothills, and backroads. Good elementary, middle & high schools. Very near Evergreen Valley College. Reasonable grocery shopping. Good restaurants just starting to emerge. "

"great neighborhood, good schools, very safe and upscale area. walking distance to grocery stores, Ranch Golf Club on the premise."

	With landslide 95138	Without landslide 95135
count	104	116
mean	$1.95M	$1.39M
min	$780000	$450000
25%	$1.53M	$1.14M
50%	$1.92 M	$1.34M
75%	$2.33M	$1.8M
max	$4.9 M	$2.32M

What is the effect of natural hazard (Fault zone) in price?

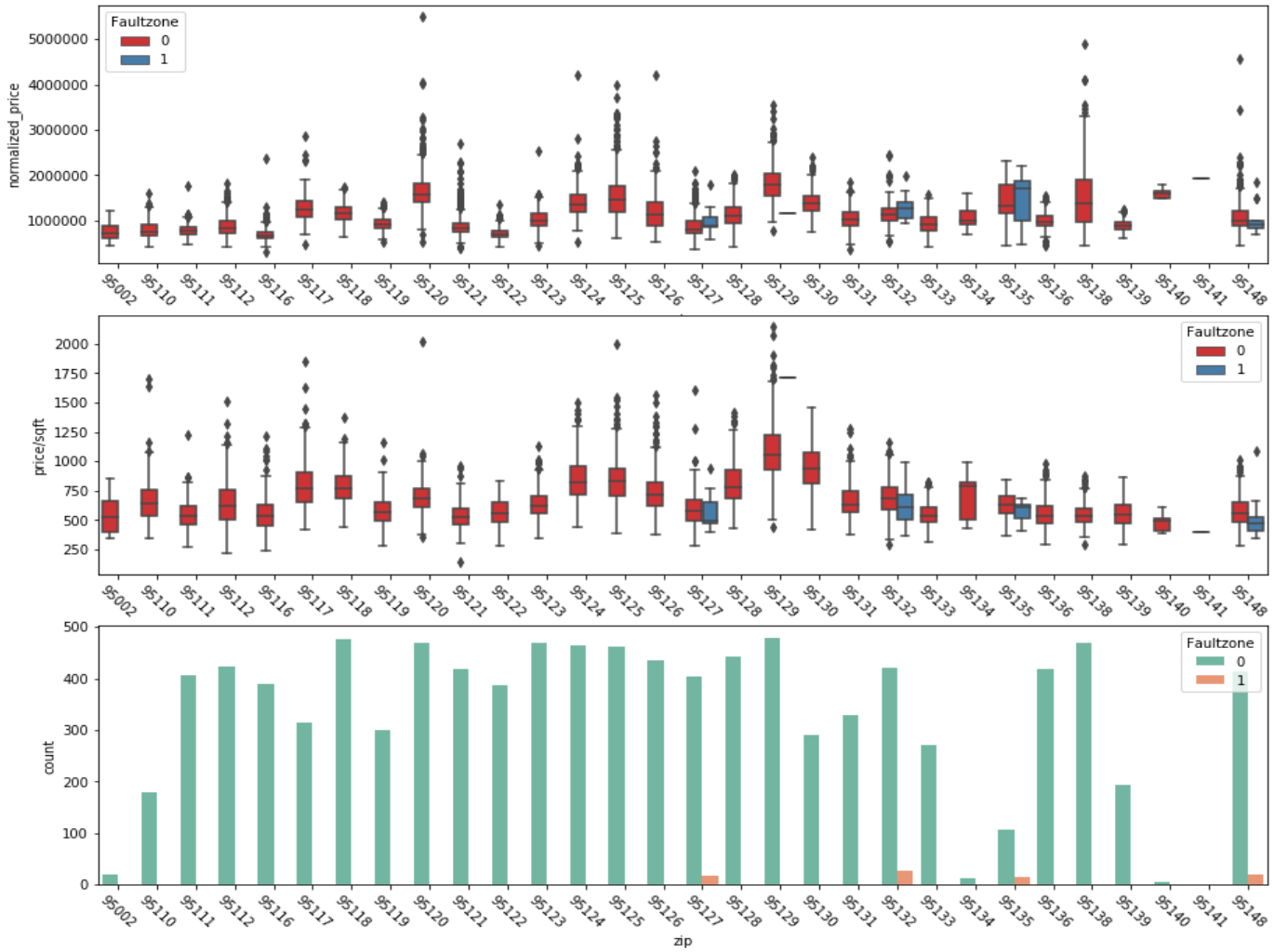


Fig 21a: Price distribution box plot (single family)

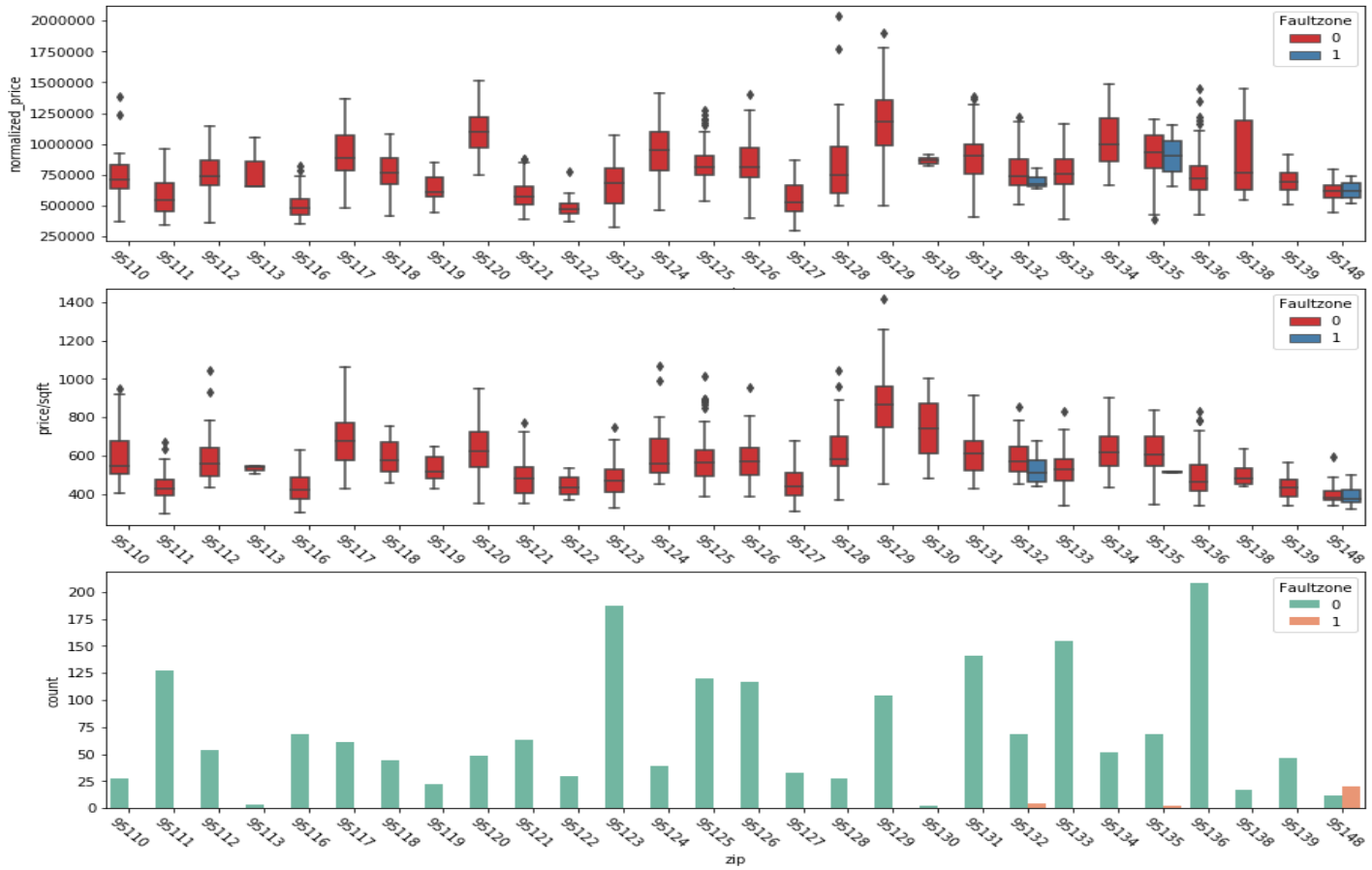


Fig 21b: Price distribution box plot (townhomes)

Above plot shows price distribution, price/sqft, count for hazard and non hazard various zip codes in single family homes. In order to study the effect of fault zone on price, equal number of hazard and non hazard homes were analyzed in nearby zip codes with equal price /sqft. For this, zip code 95127 with fault hazard and 95140 with no fault hazard were considered. Below median value table indicates the fault hazard prone homes are selling less than non fault hazard areas.

	With fault zone 95127	Without fault zone 95140
count	17	5
mean	$980327	$1.6 M
min	$575000	$1.5 M
25%	$846000	$1.5 M
50%	$884000	$1.59 M
75%	$1.07 M	$1.65 M
max	$1.8 M	$1.78 M

What is the effect of natural hazard (Fault zone) in price?

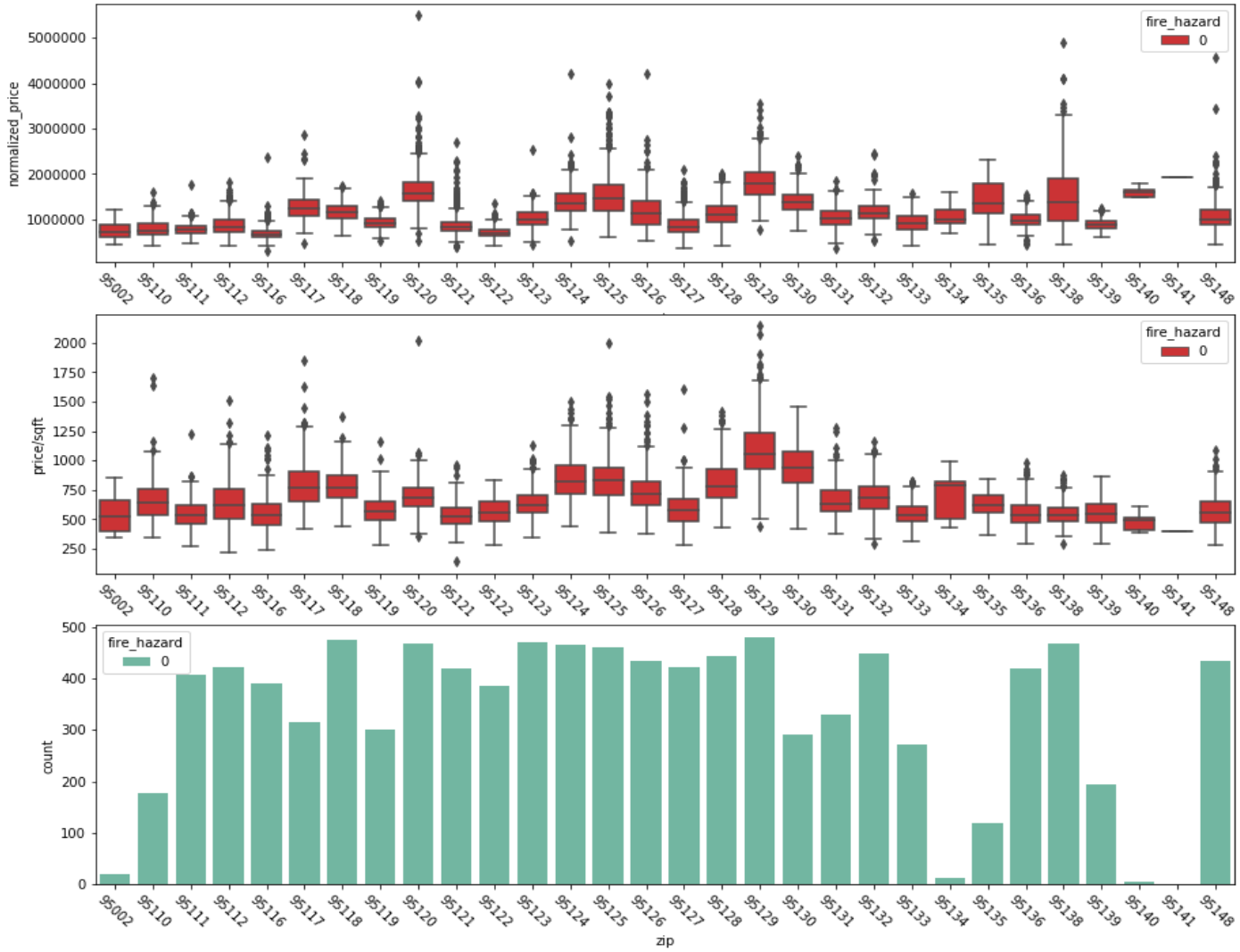


Fig 22a: Price distribution box plot (single family)

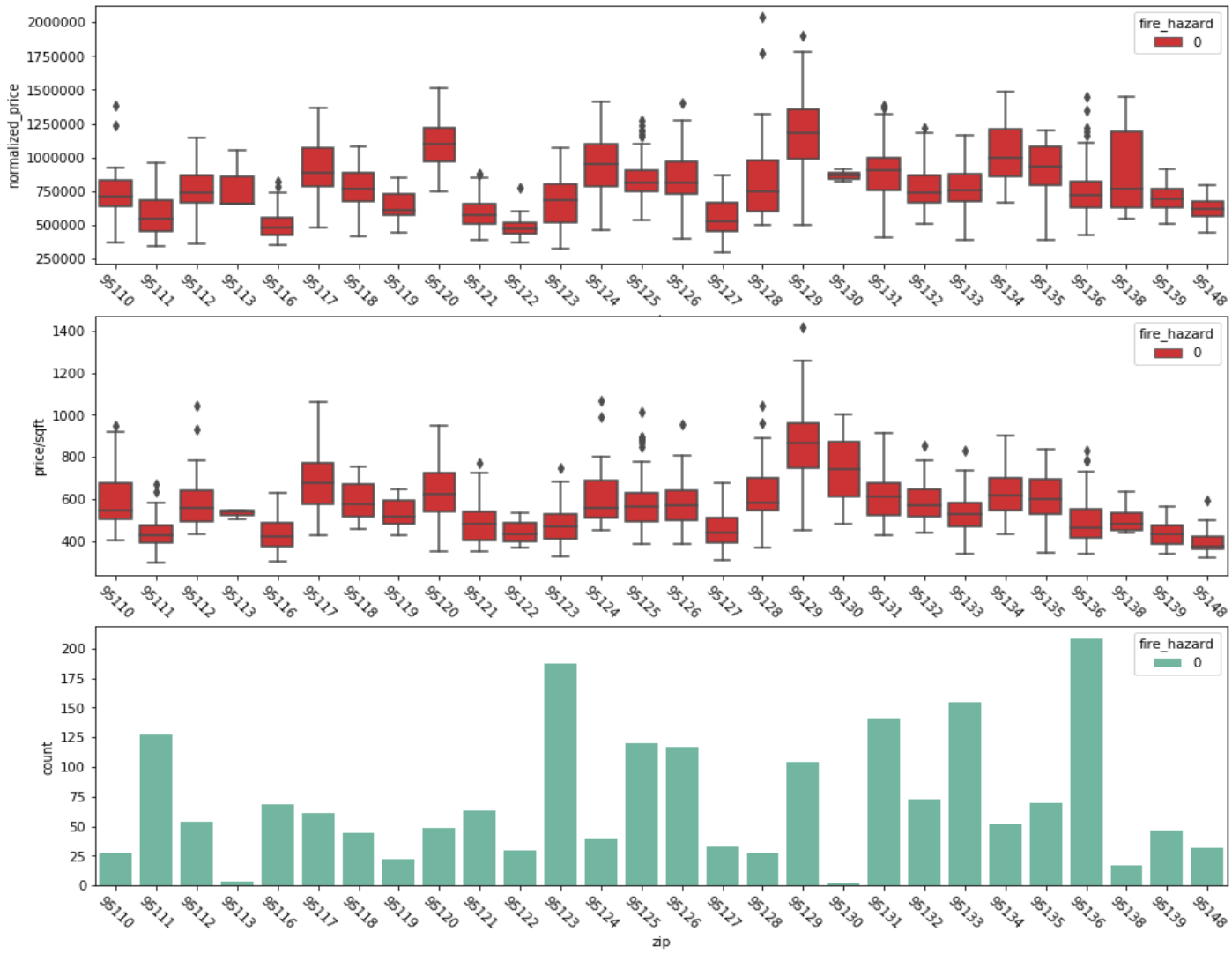


Fig 22b: Price distribution box plot (townhomes)

There is no fire hazard in sold properties in all zip codes.

Scatter plot was plotted to see distribution of sqft vs price for single family and townhomes.

What is the distribution of price per zip code?

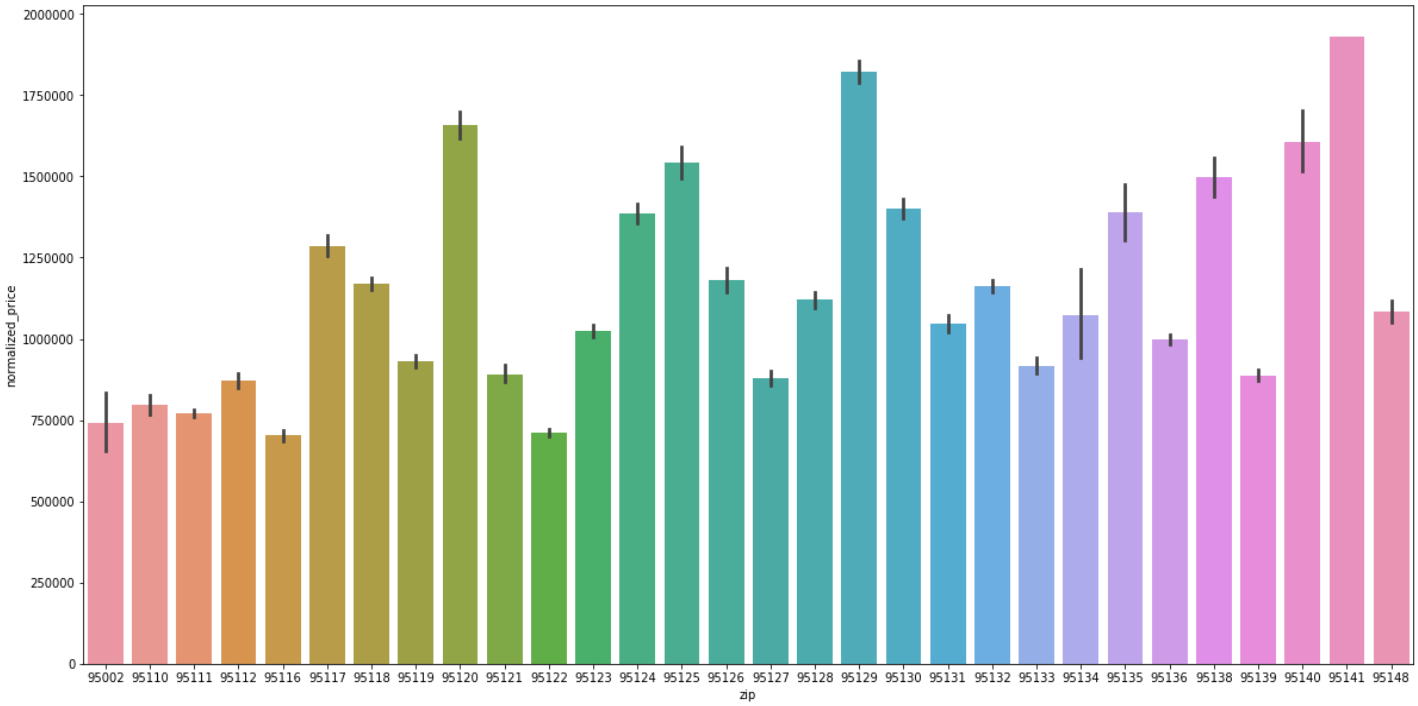


Fig 23a: Price distribution bar plot (single family)

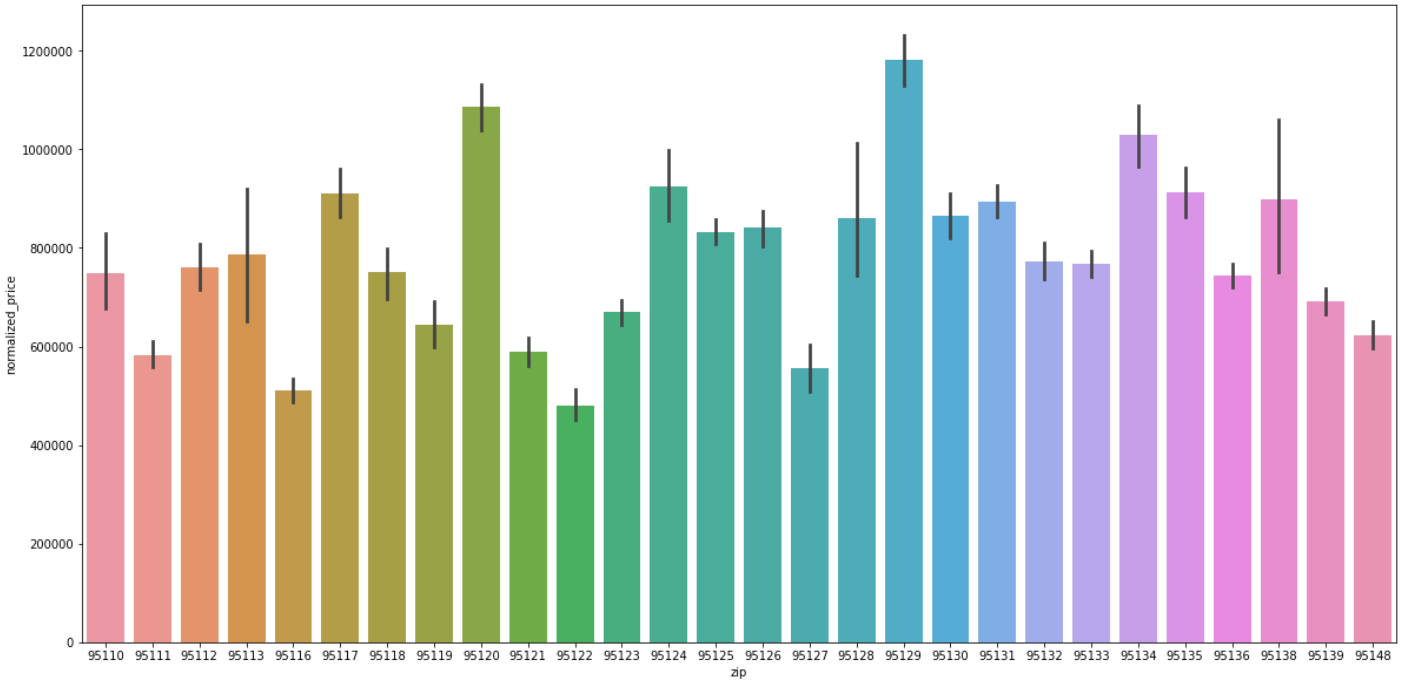


Fig 23a: Price distribution bar plot (townhomes)

What is the distribution of price per year built?

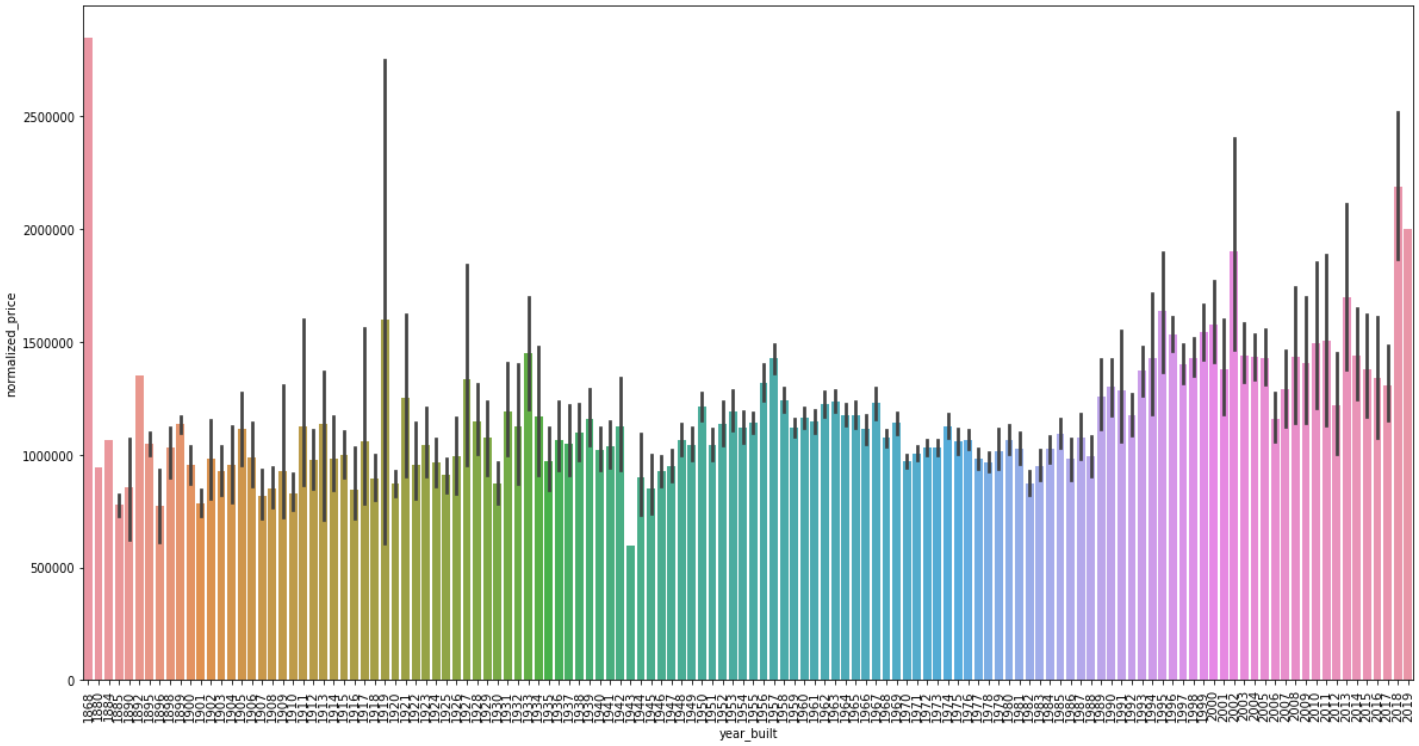


Fig 24a: Price distribution bar plot (single family)

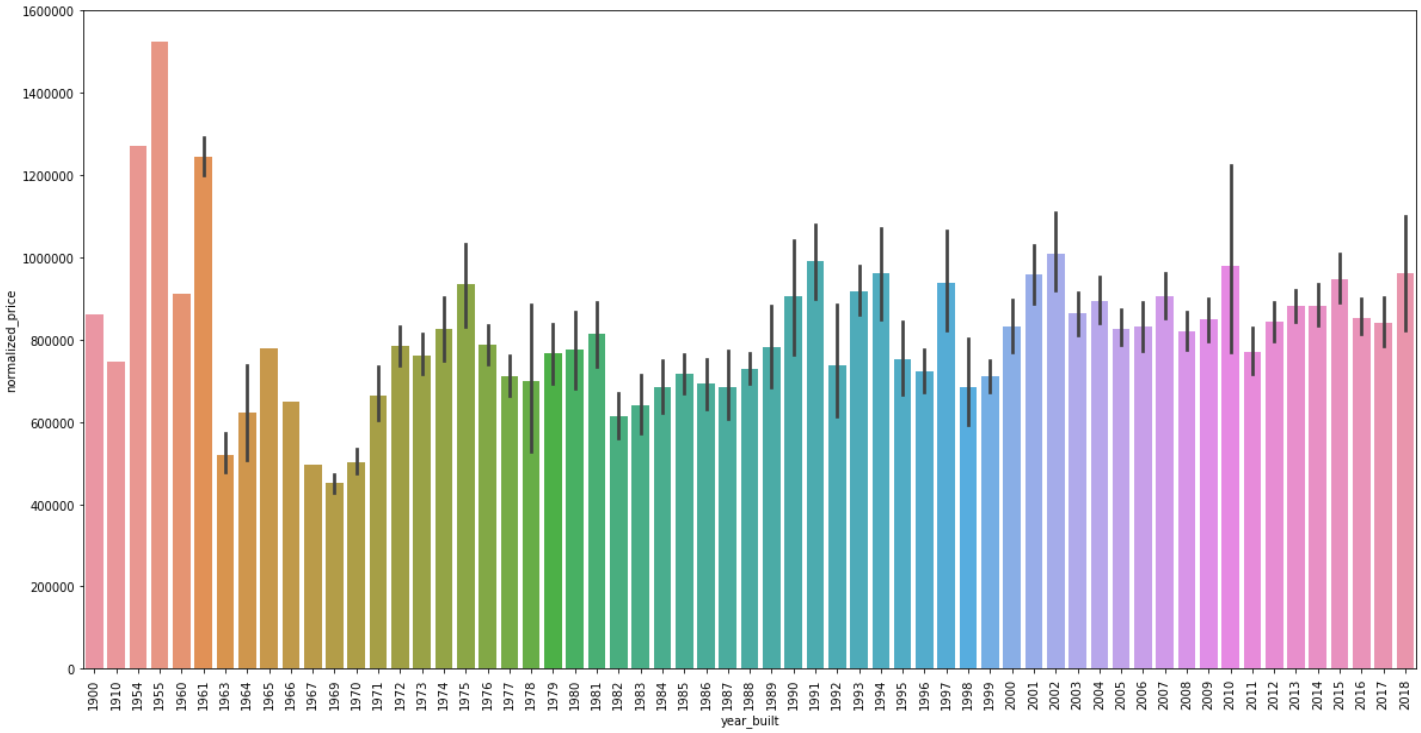
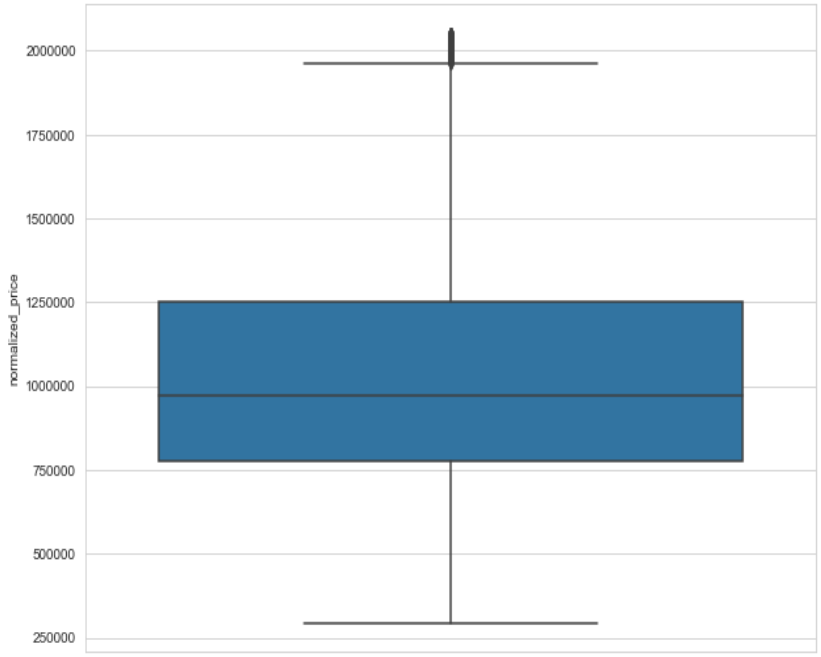
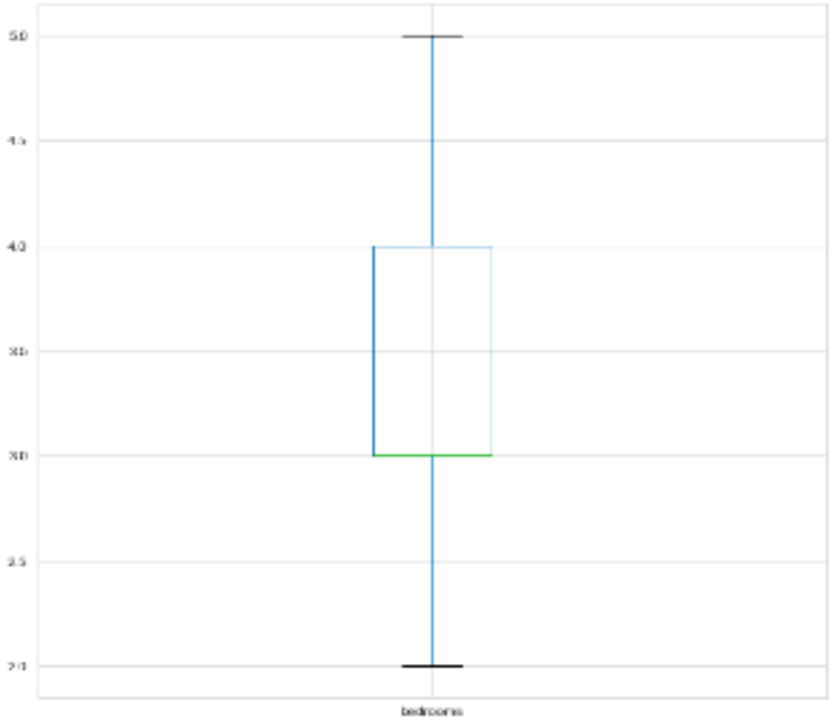


Fig 24b: Price distribution bar plot (townhomes)

Bar plot distribution for all features





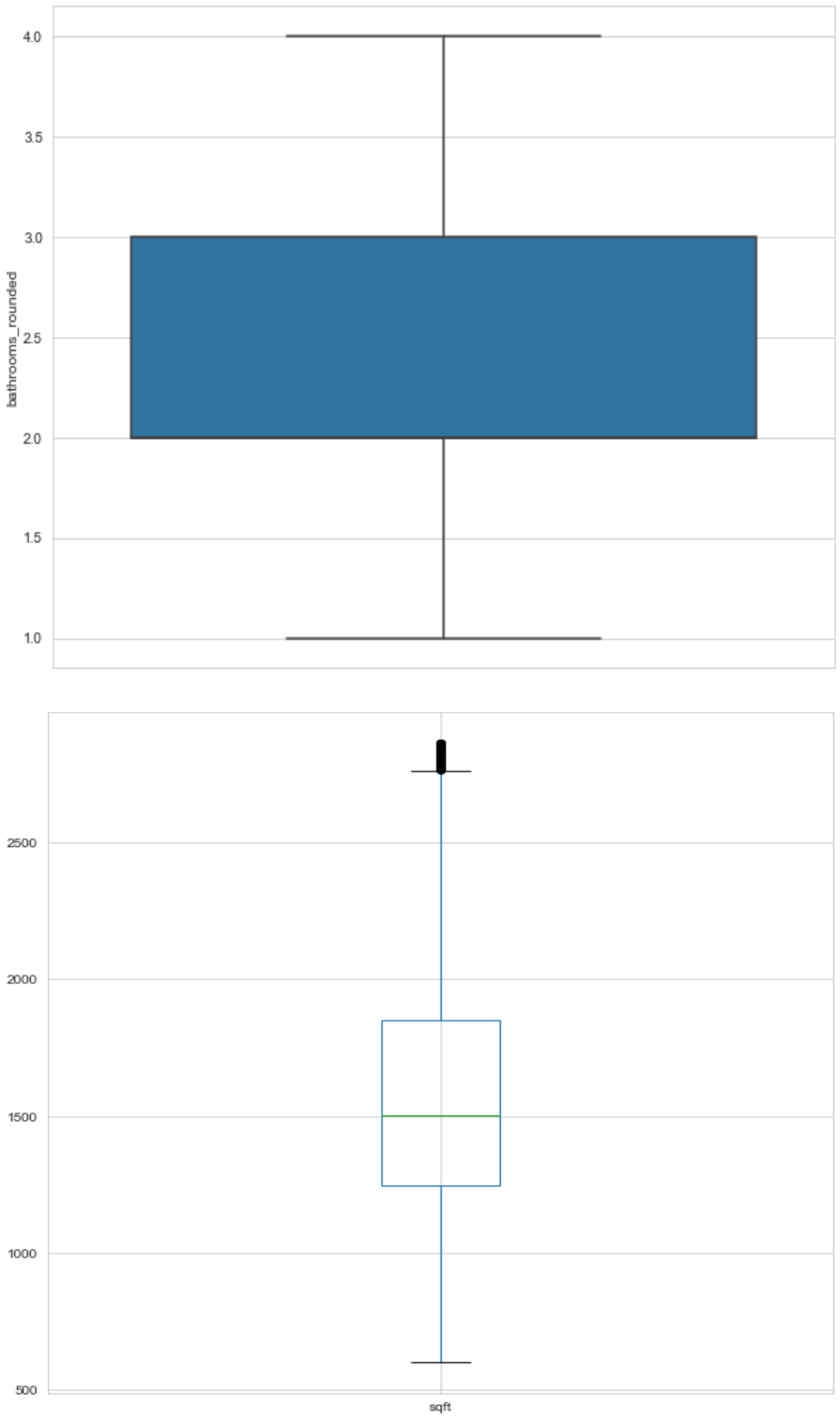


Fig 25: Bar plot distribution

Price distribution in geospatial frame

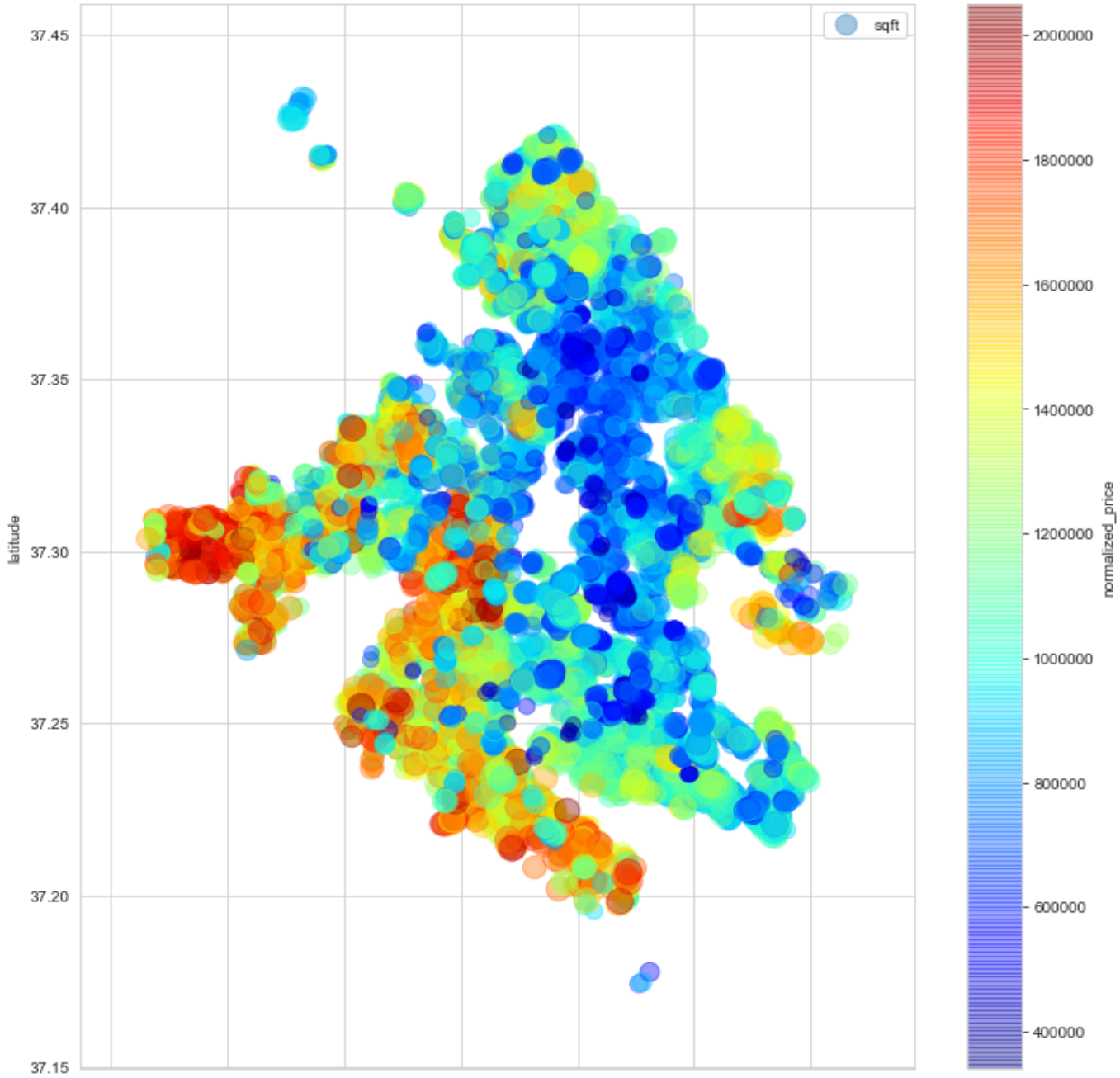
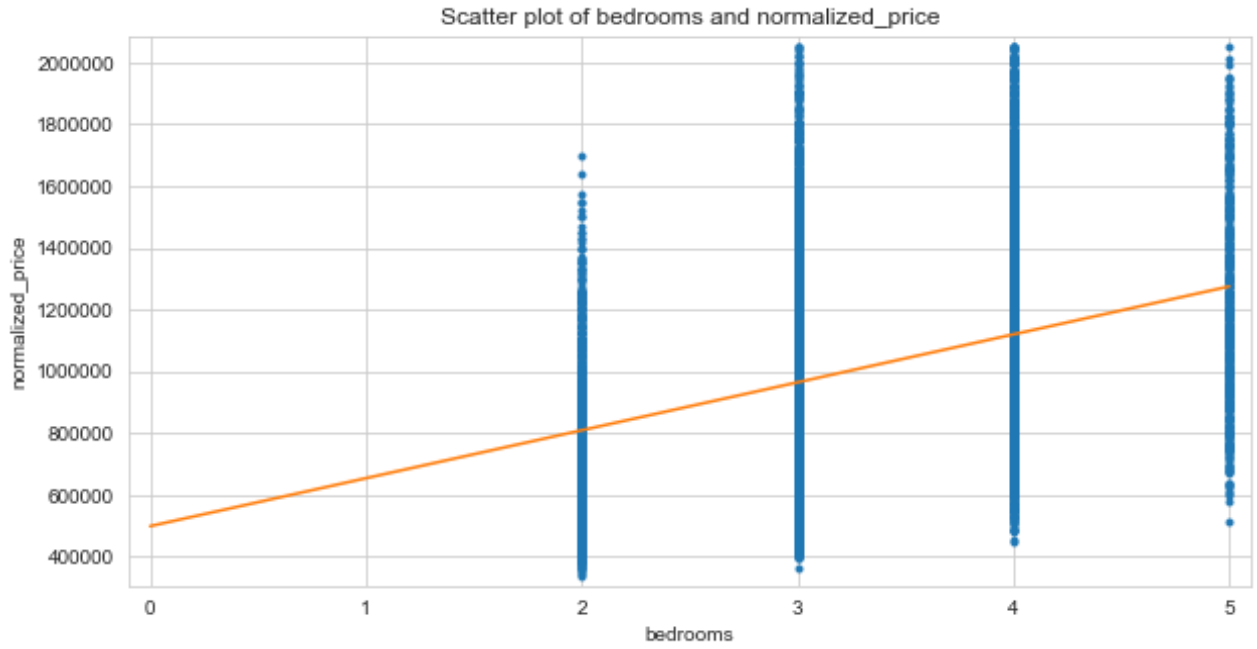


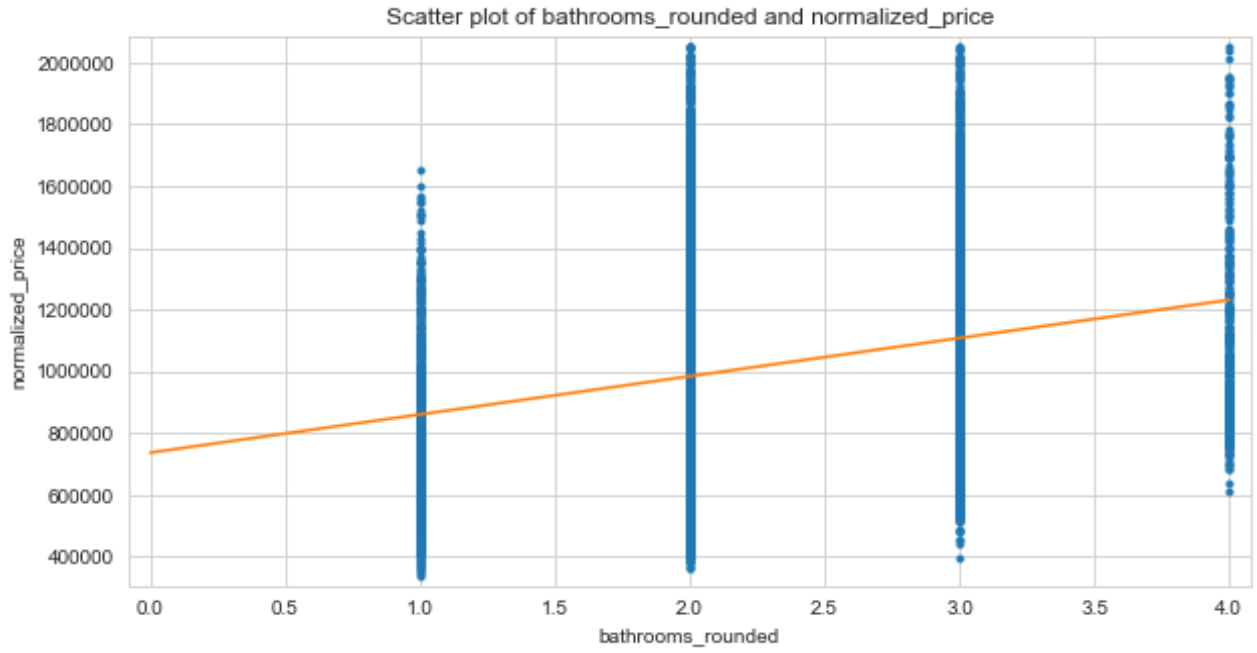
Fig 26: price distribution in geospatial frame

Above plot shows high price in west and south side of san jose. Extreme south is the in highest price due to close proximity to tech companies in sunnyvale and Santa Clara. Above generated plot is inline with price market trend plot observed by Trulia.

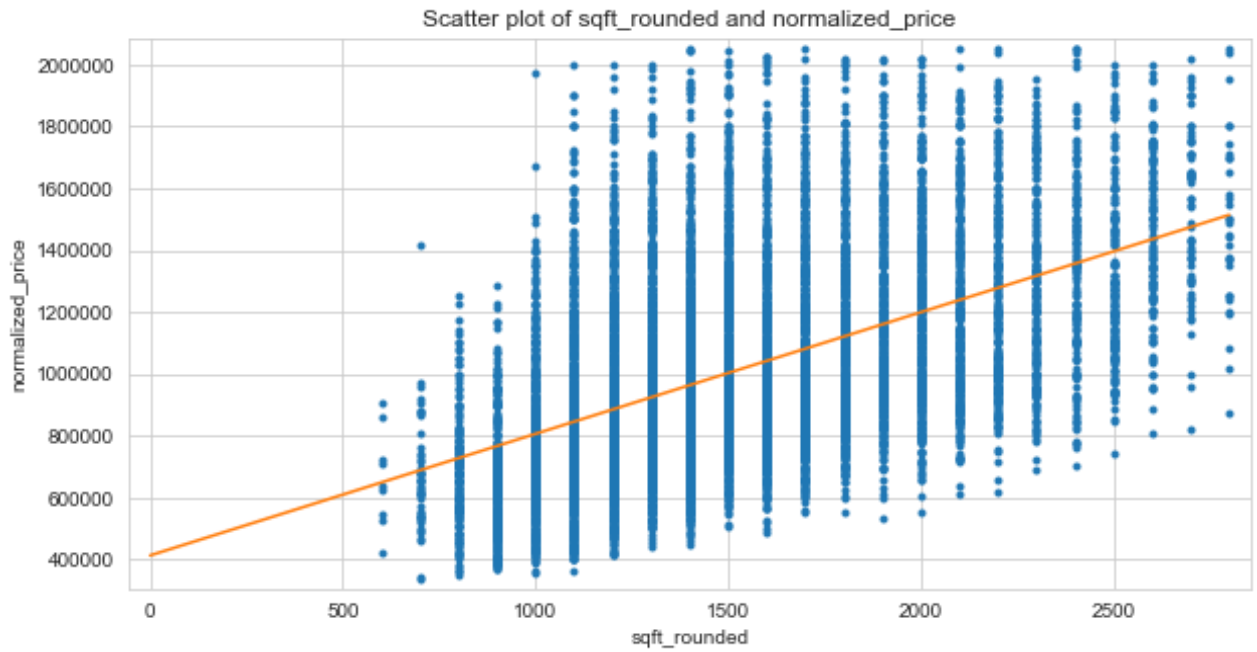
Scatter plot distribution



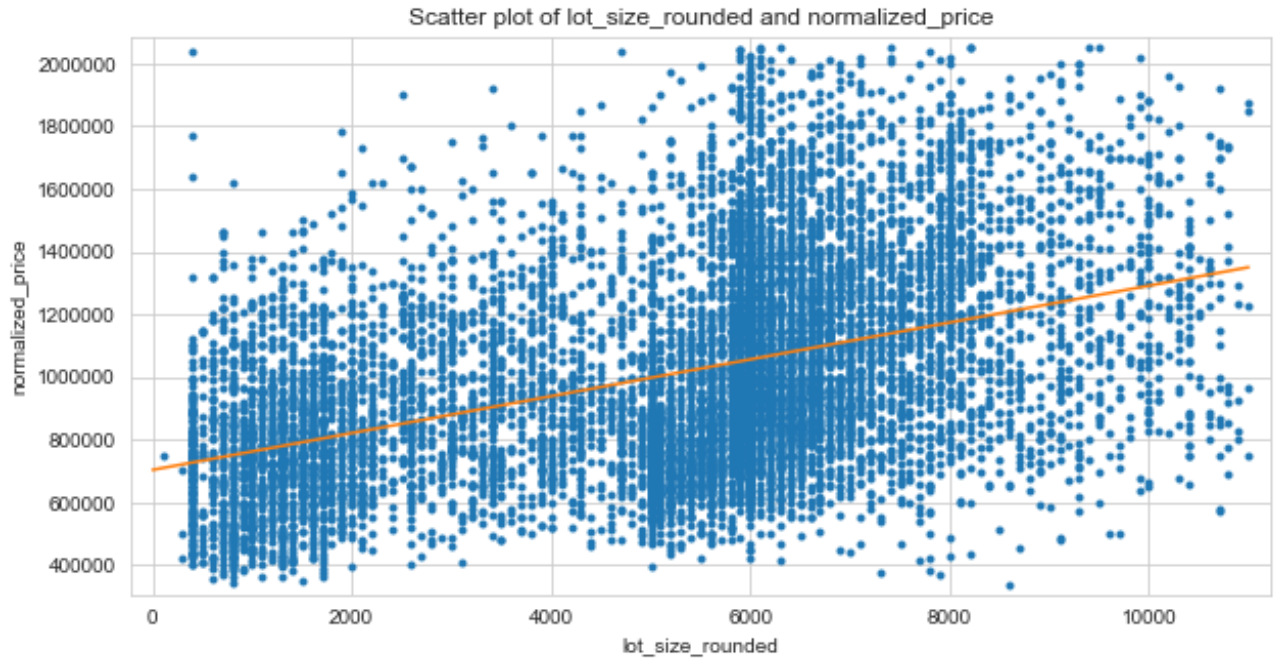
Number of bedrooms shows positive correlation with price



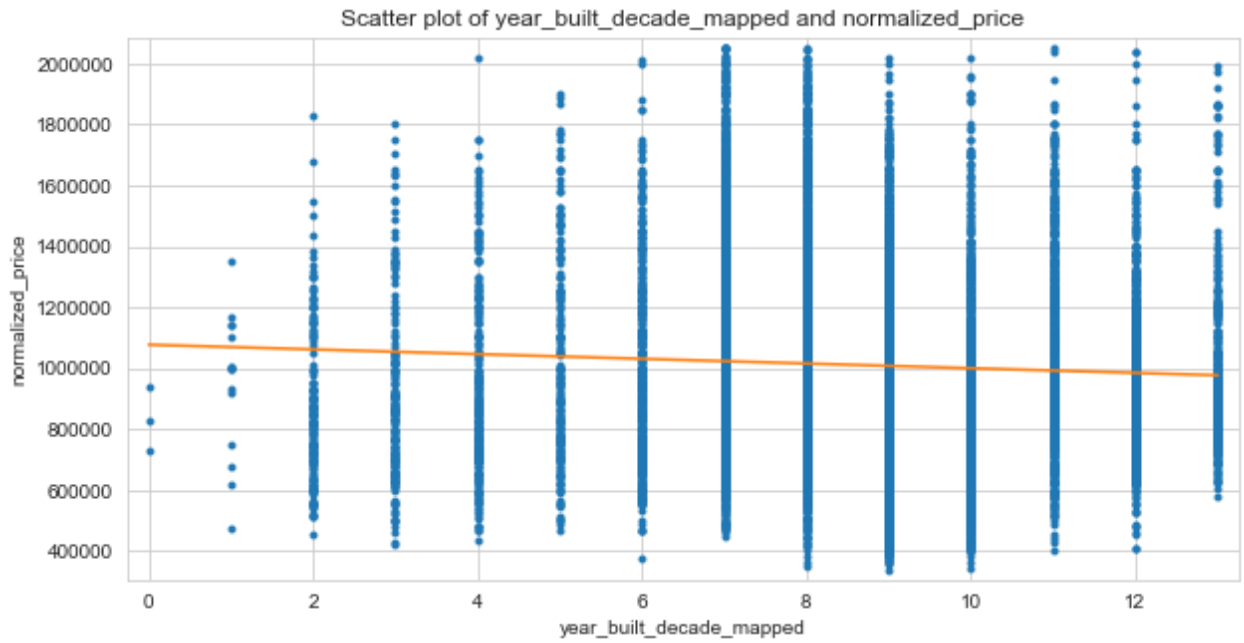
Number of bathrooms shows positive correlation with price



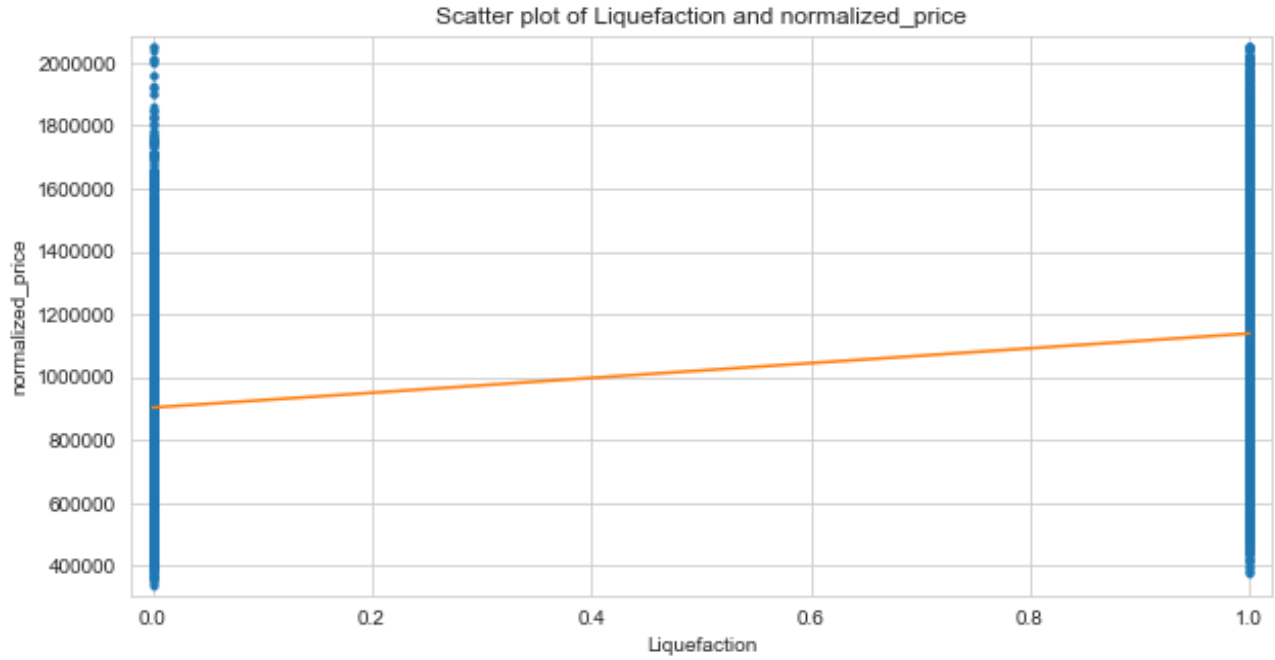
sqft shows positive correlation with price.



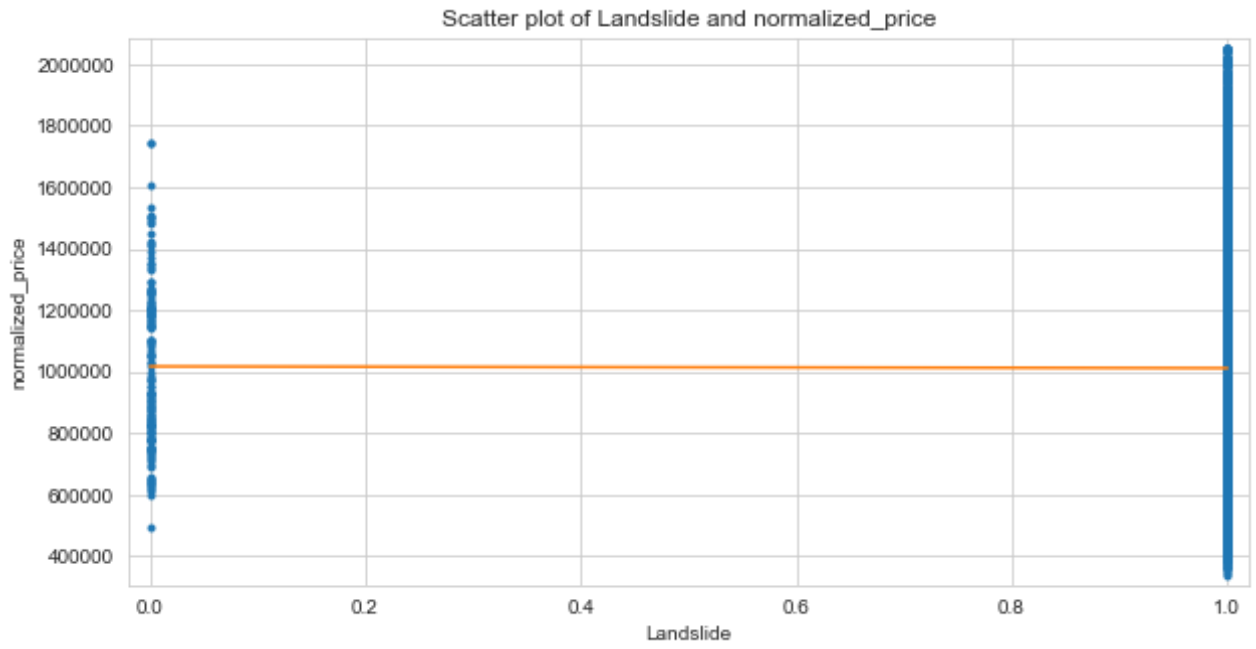
Lot size shows positive correlation with price



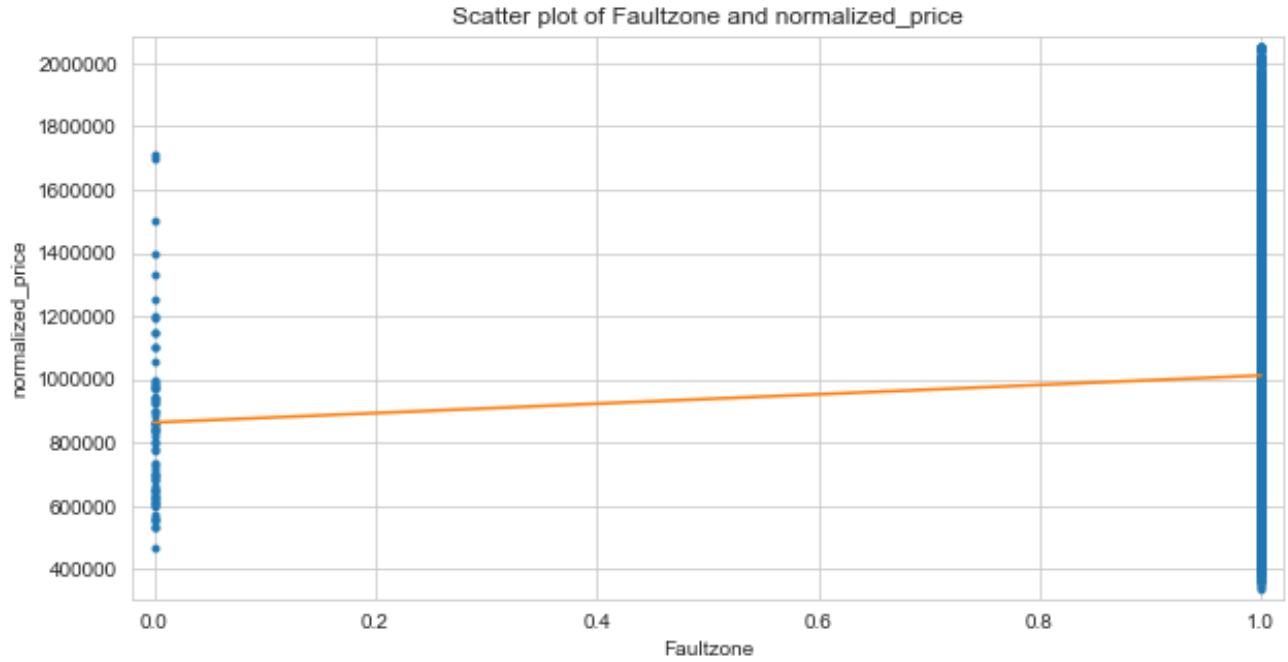
Year built shows weak negative correlation with price



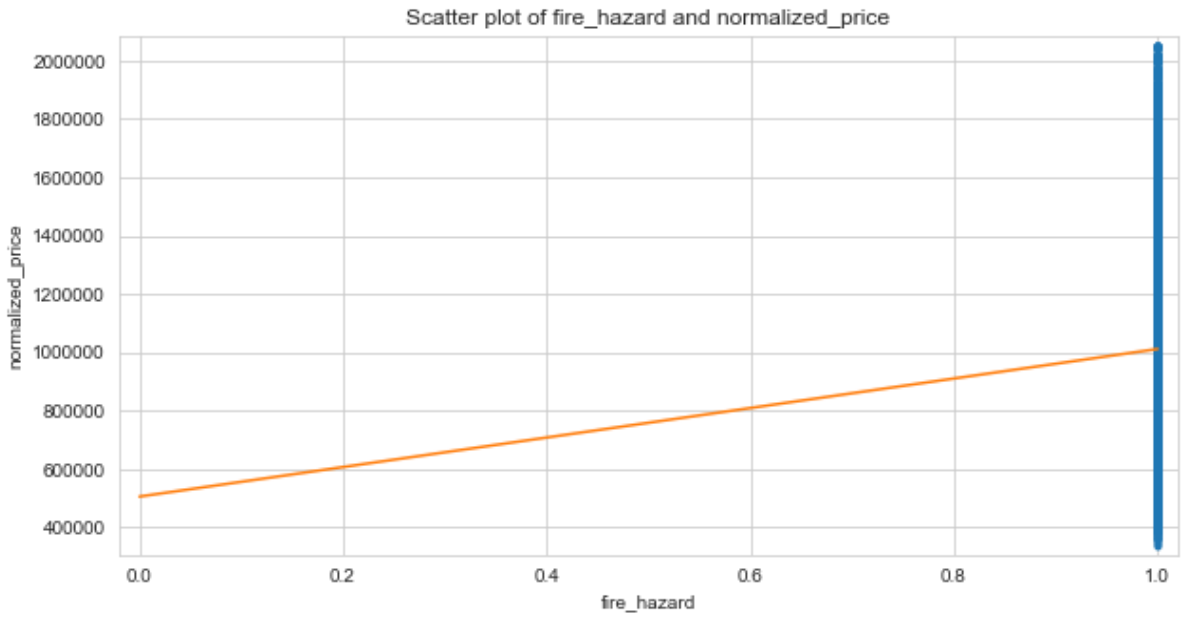
Liquefaction shows positive correlation with price



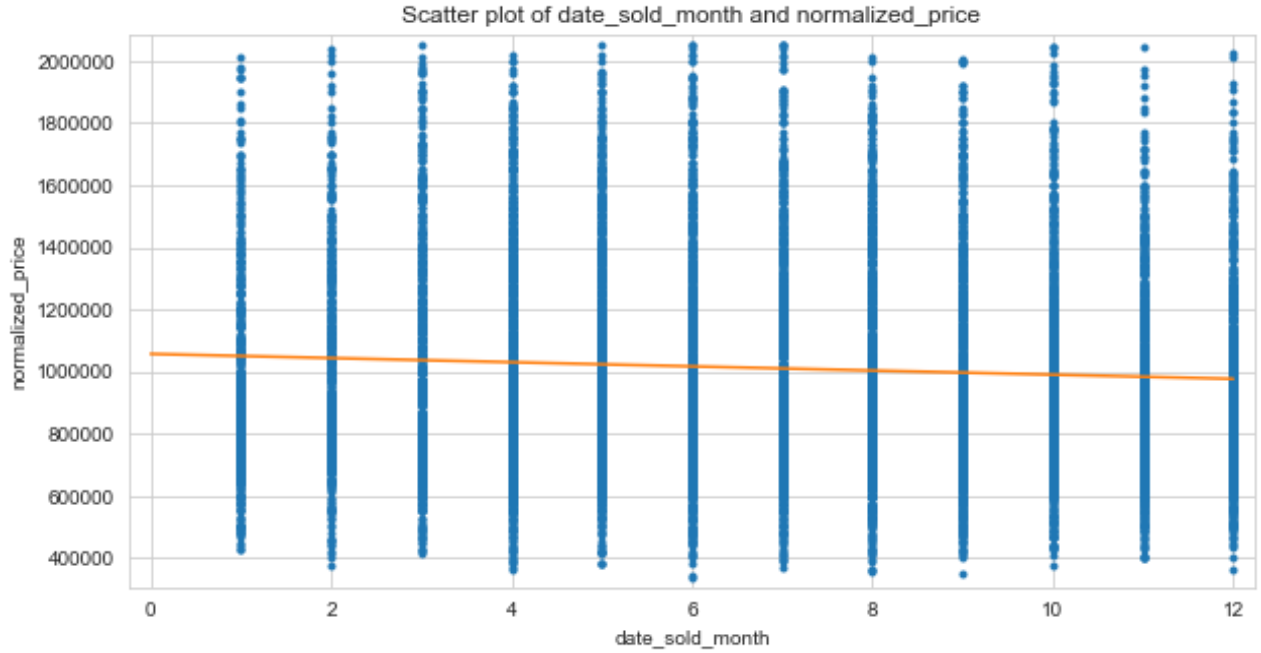
Landslide shows weak negative correlation with price



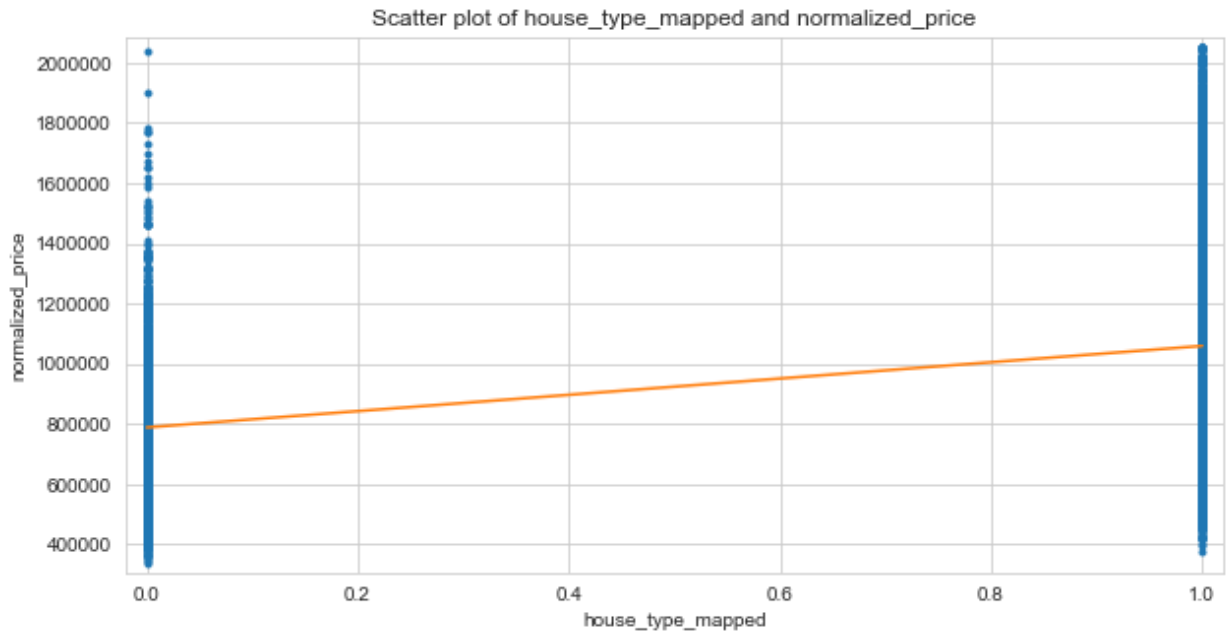
Fault zone shows positive correlation with price



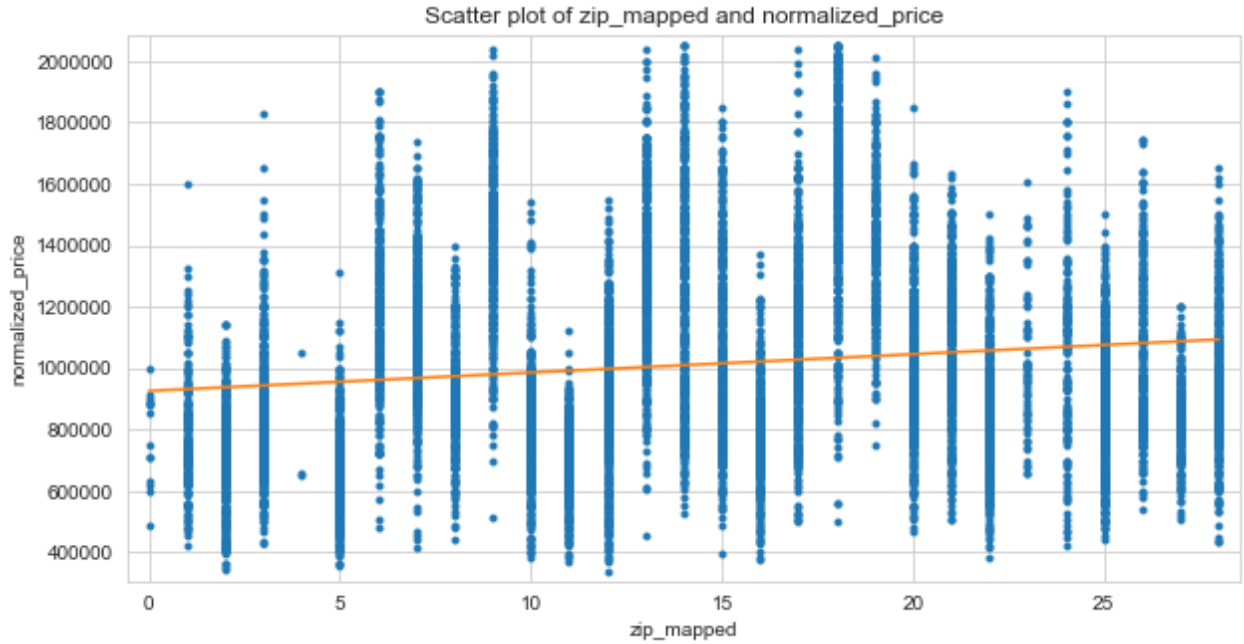
No correlation between fire hazard and price



Sold month shows negative correlation with price

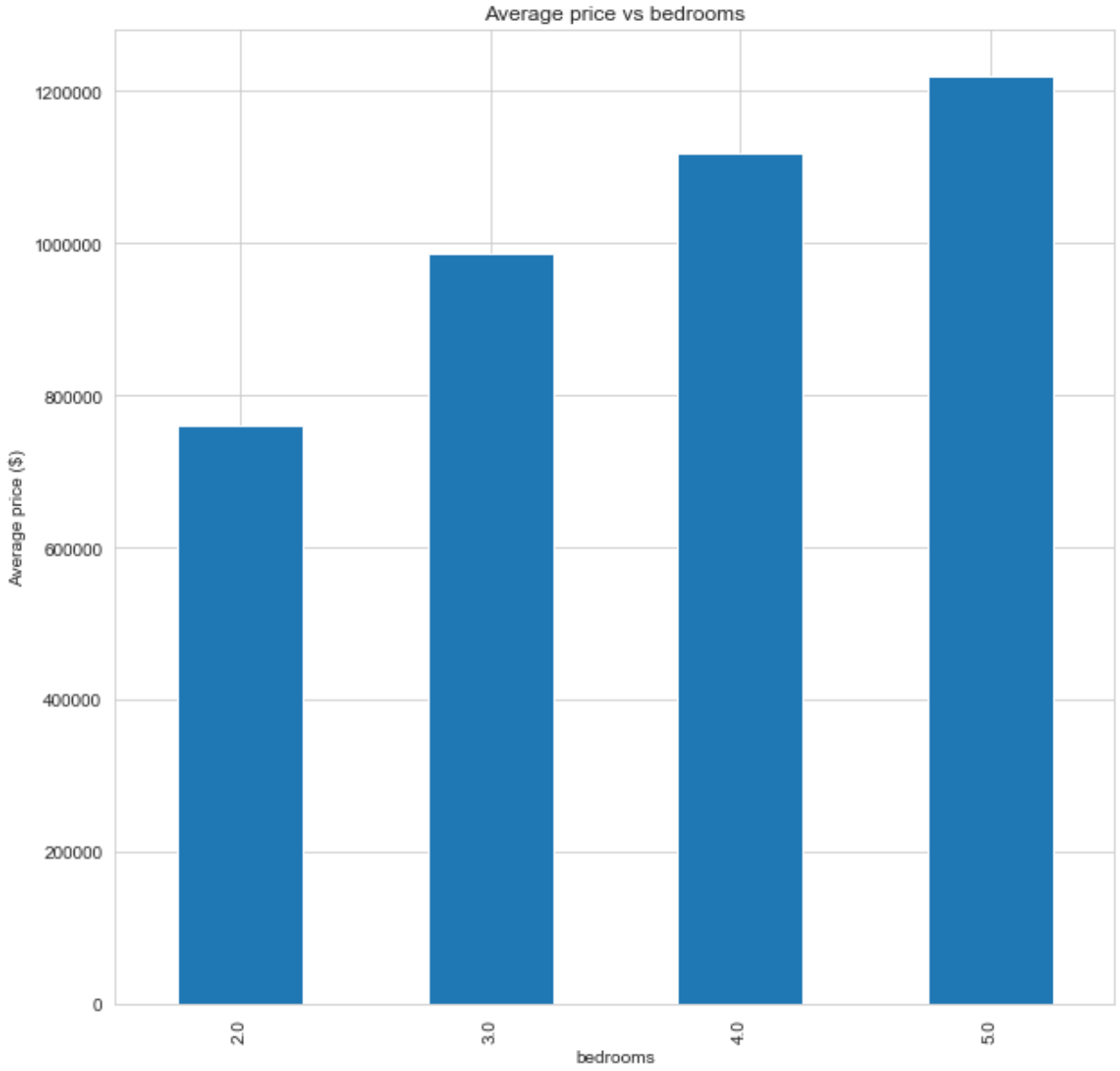


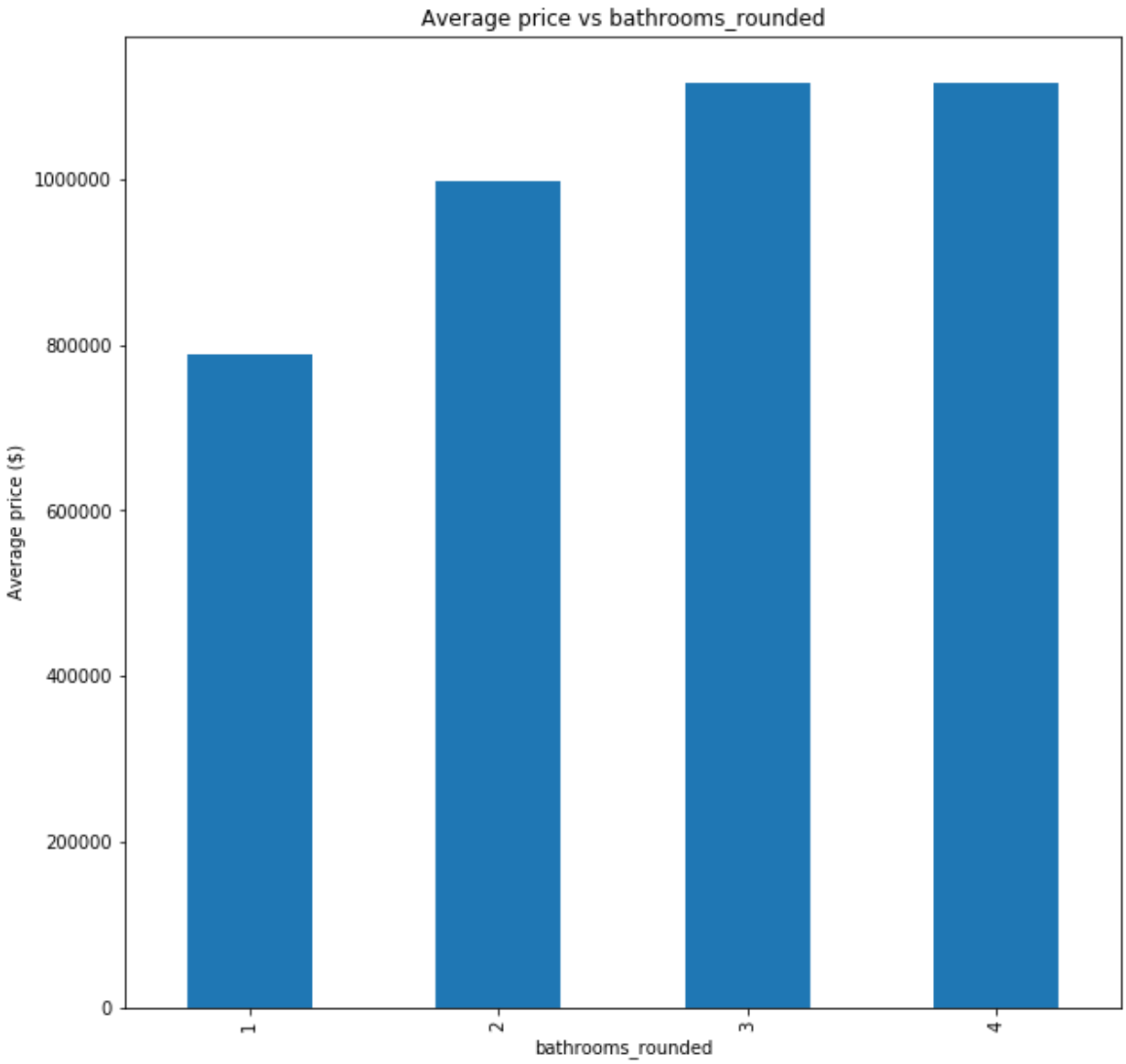
House type shows positive correlation with price

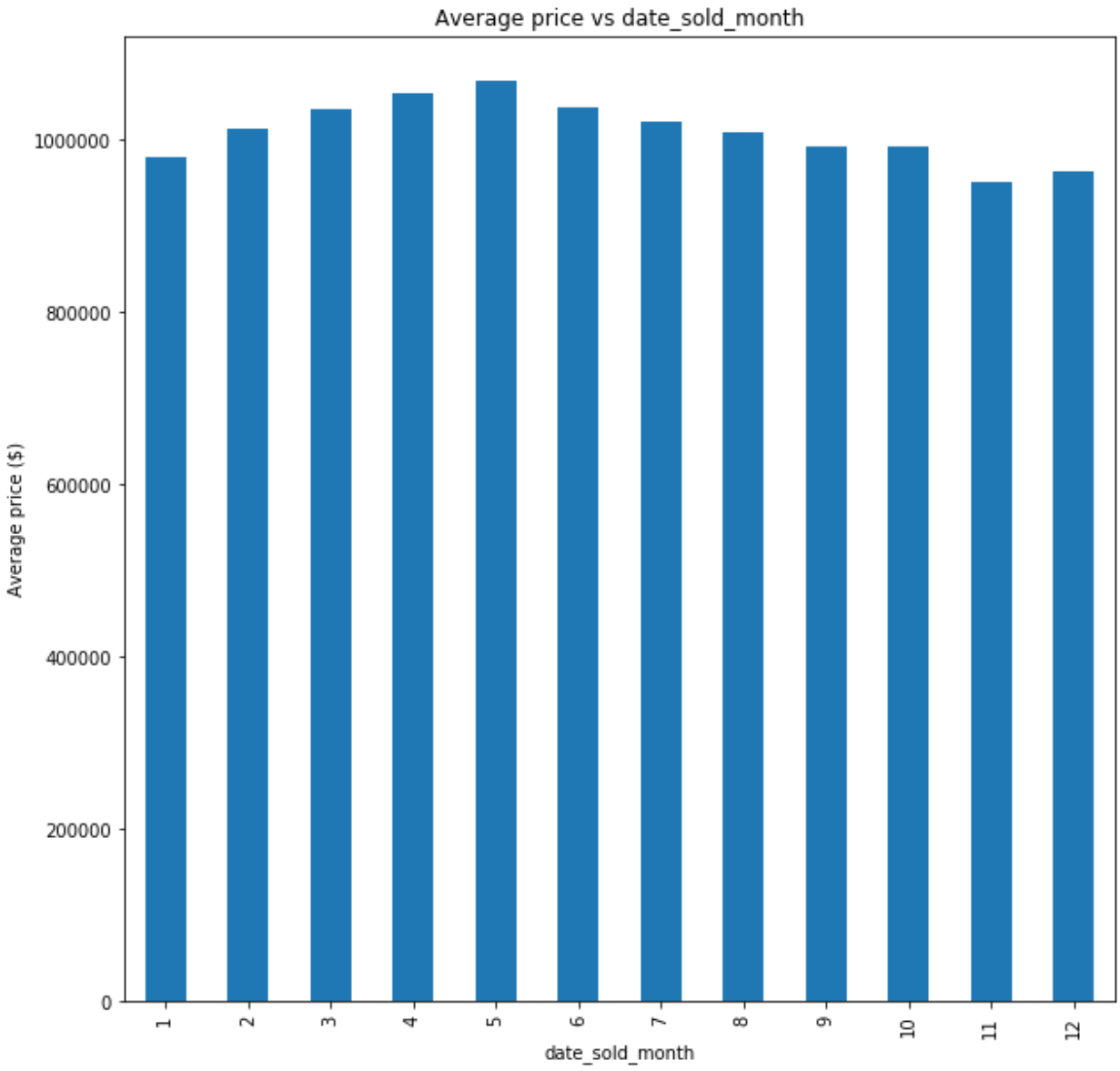


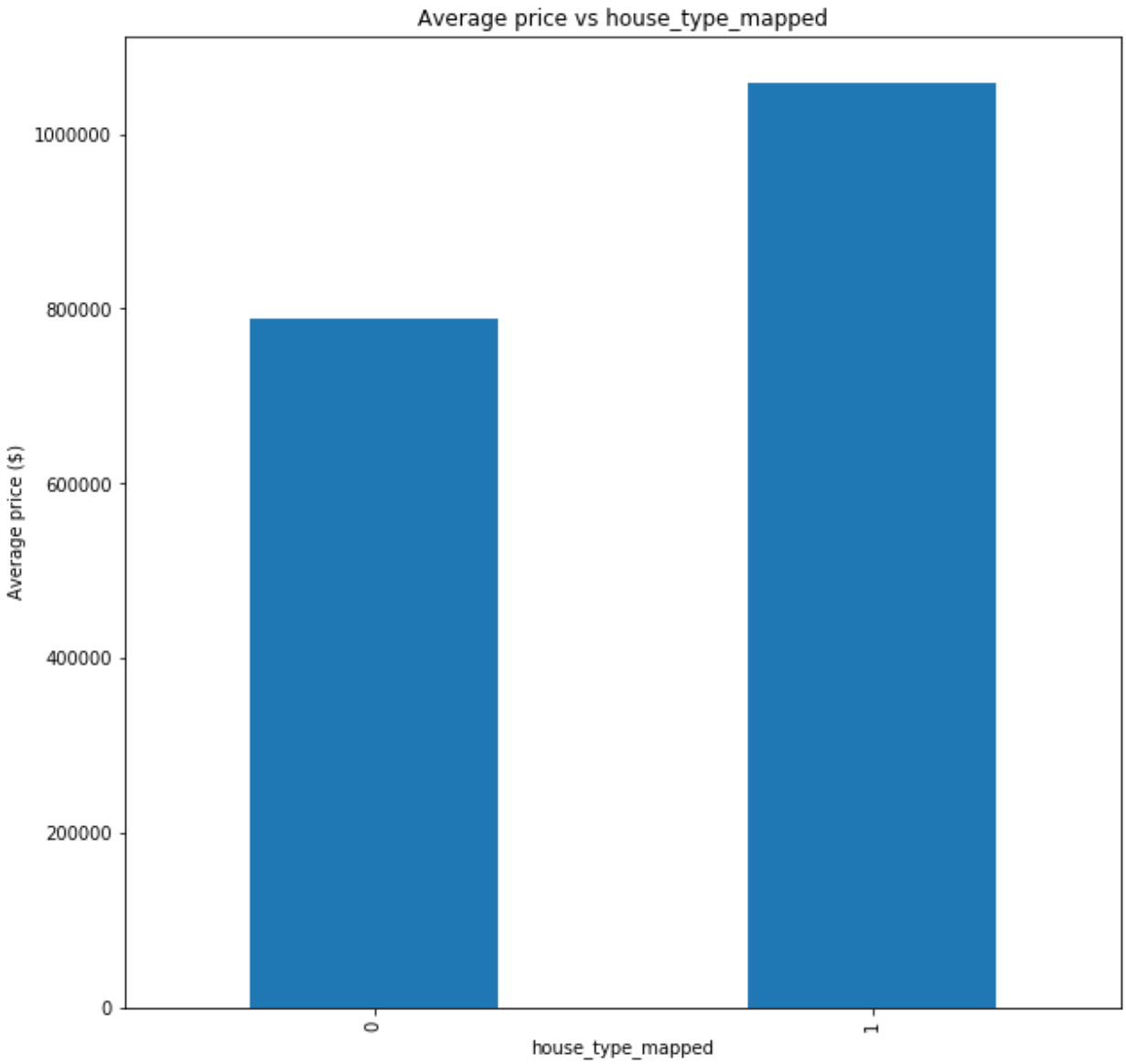
Zip code shows positive correlation with price

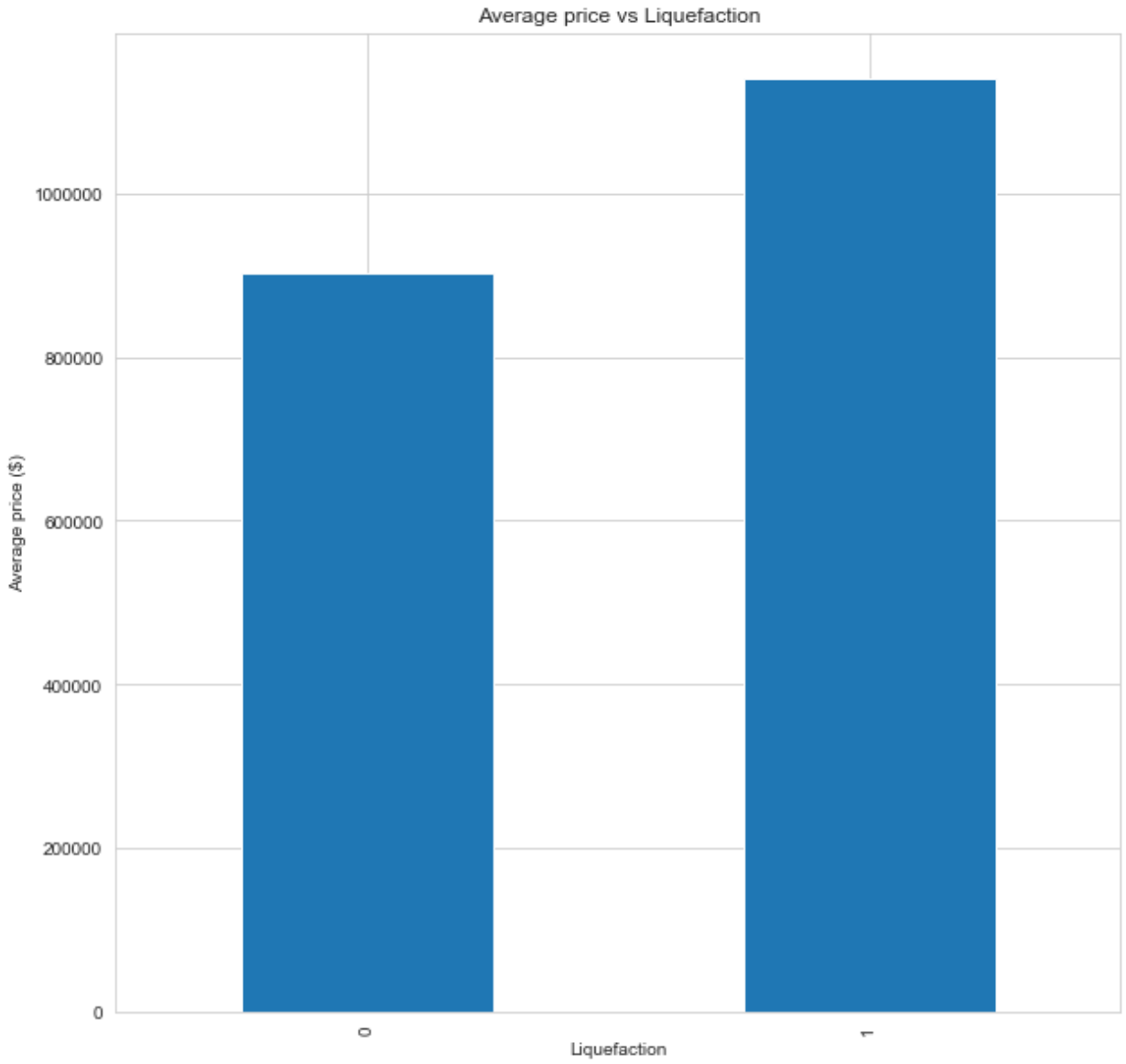
Average price bar plot distribution

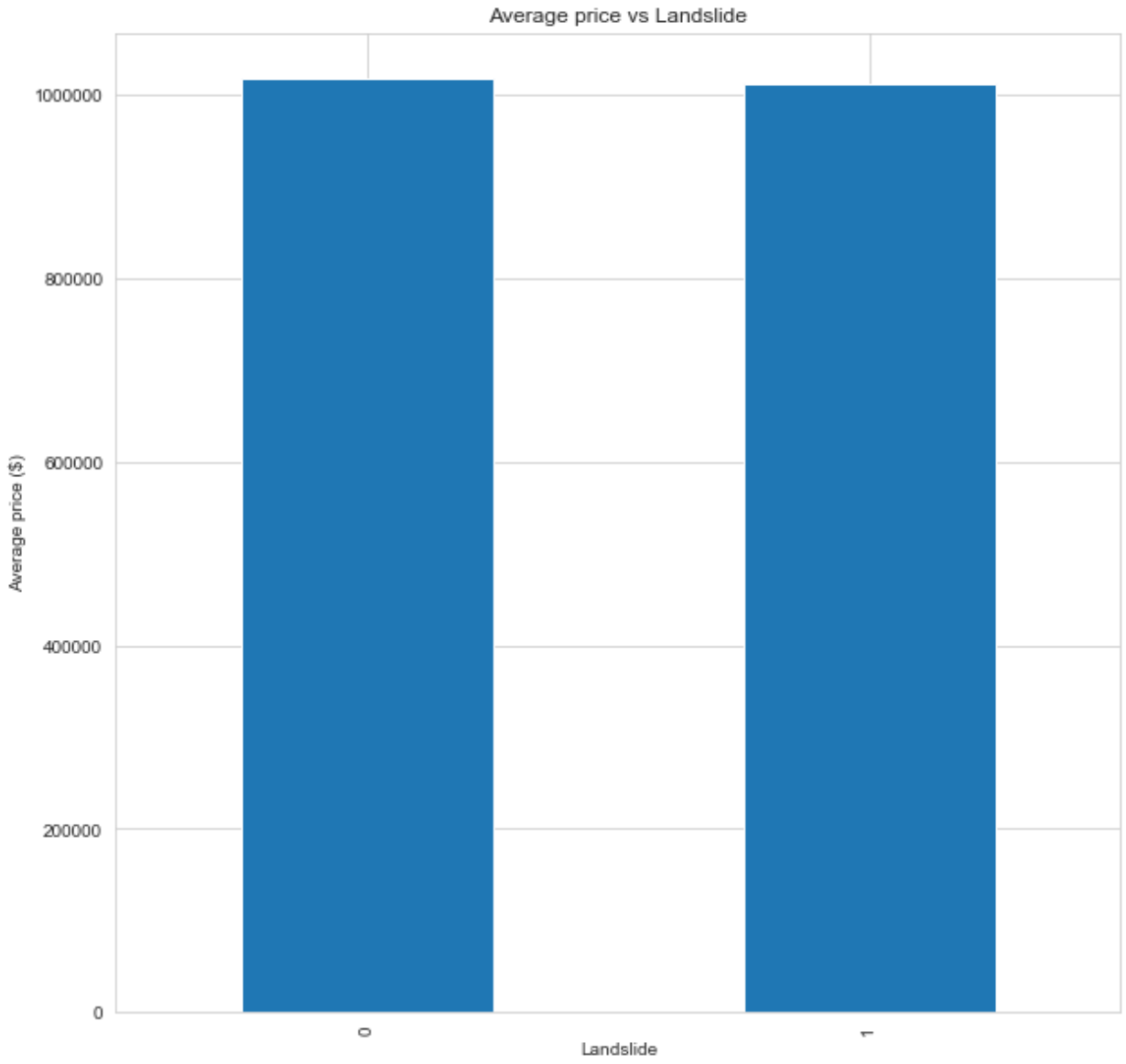


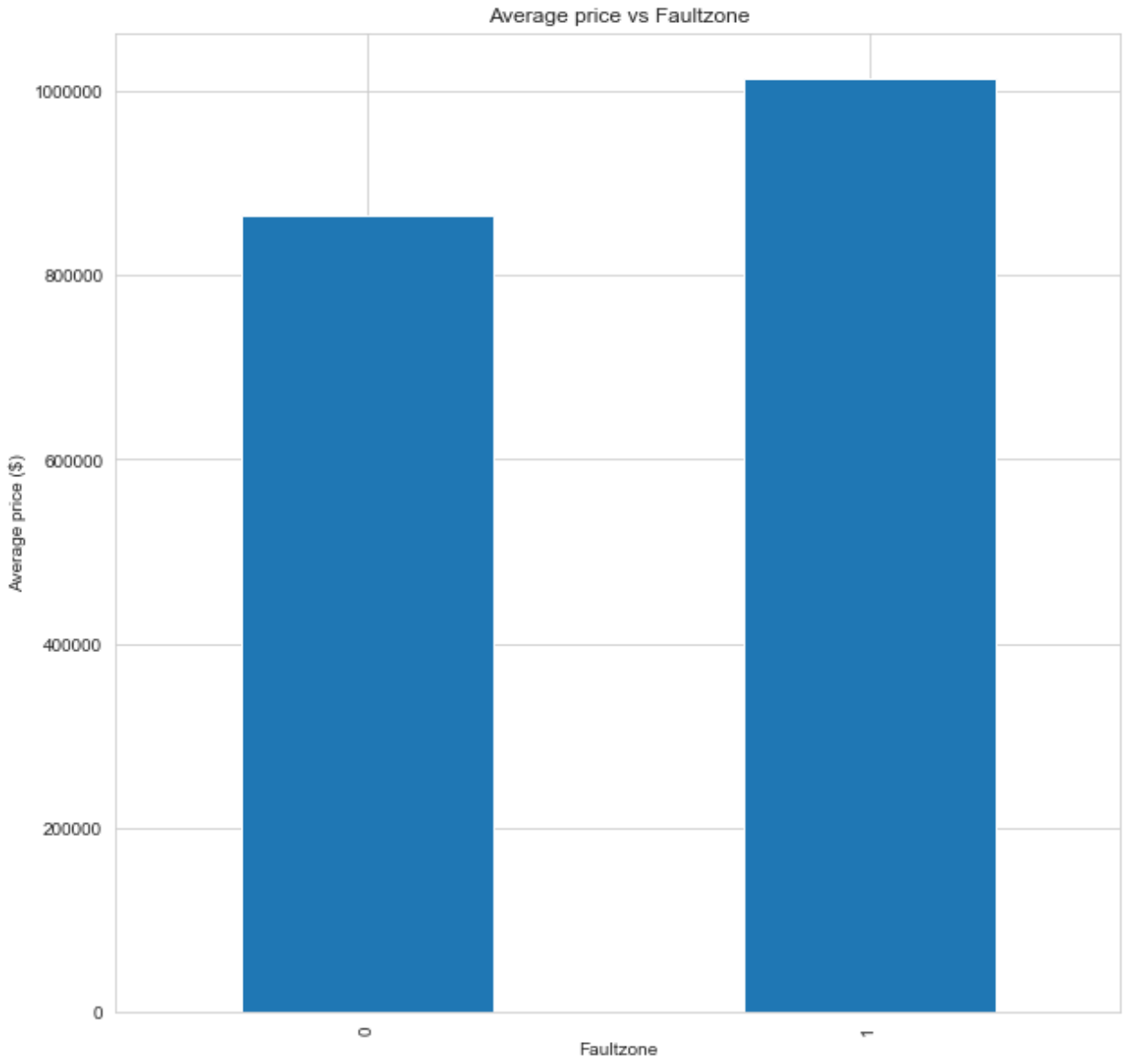

Average price vs bathrooms_rounded
Average price ($)
0
200000
400000
600000
800000
1000000
1
2
3
4
bathrooms_rounded


Average price vs date_sold_month
Average price ($)
0
200000
400000
600000
800000
1000000
1
2
3
4
5
6
7
8
9
10
11
12
date_sold_month


Average price vs house_type_mapped
1000000
800000
600000
400000
200000
0
Average price ($)
0
1
house_type_mapped


Average price vs Liquefaction
1000000
800000
600000
400000
200000
0
Average price ($)
0
1
Liquefaction


Average price vs Landslide
1000000
800000
600000
400000
200000
0
Average price ($)
0
1
Landslide


Average price vs Faultzone
1000000
800000
600000
400000
200000
0
Average price ($)
0
1
Faultzone

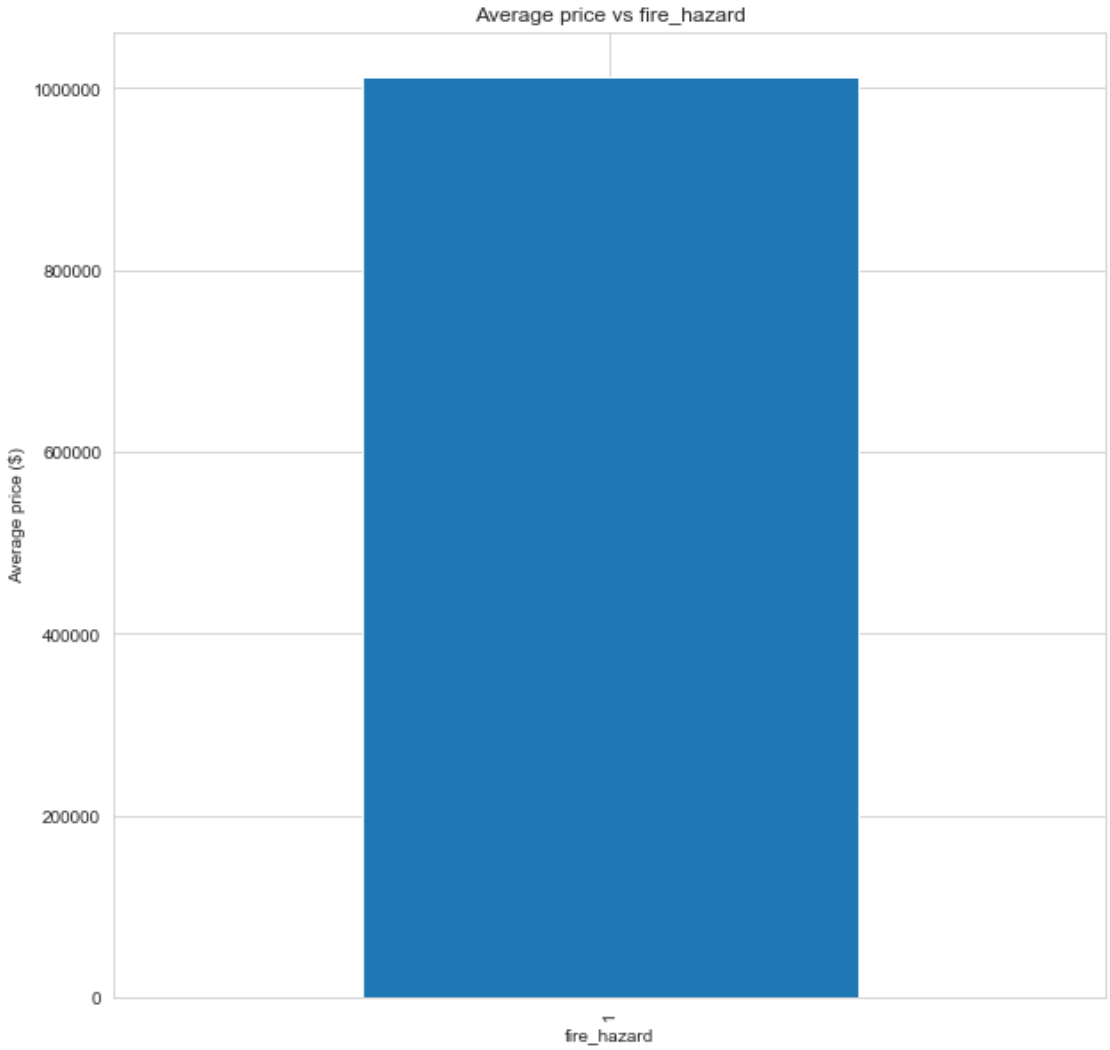


Fig 27: Average price distribution

Inferential Statistics

A hypothesis test was conducted to check if there is a significant correlation between year_built, liquefaction hazard, landslide hazard, sqft, fault zone hazard, zip code, bedrooms, bathrooms, house_type, lot size and price. For theses tests, the null hypothesis was assigned with no significant correlation and alternative hypothesis with significant correlation between features. If the p-value for the hypothesis test is less than the level of significance 0.05, then the null hypothesis will be rejected which suggests there is a correlation. p value less than 0.05 was obtained for bedrooms, bathrooms, sqft, lot size, liquefaction, fault zone, house type and zip code which suggested there was a correlation between those features and price. Other features such as year built, sold month, landslide, p value was greater than 0.05 which suggested there was no significant correlation between those features and price.

Appendix 1

Table 1: Zip codes vs Neighborhood

Zip codes	Neighborhood
95120	Alameden valley
95127	Alum rock
95002	Alviso
95123,95136	Blossom valley
95128	Burbank
95112	Chinatown
95110,95112,95113	Downtown
95127	East foothills
95111,95123,95136	Edenvale
95148,95121,95138	Evergreen
95112	Japantown
95126	Midtown,Rosegarden
95119,95138,95139,95193,95123	Santa Teresa
95111	Seven trees
95138	Silver creek valley
95113	SOFA district
95111,95119,95120,95123,95136,95138,95139,95193	South San Jose