House Price Prediction in Natural Hazard Prone Areas

Capstone Project 1: Final Report

PROJECT PROPOSAL

1. What is the problem you want to solve?

Goal:

Given a property details and natural hazard attributes, predict the median house price in San Jose city using Zillow property data and natural hazards data such as earthquake, landslide and fire hazards. The main goal for this project is to find major attributes for the contribution of house prices in San Jose city.

Background:

California is well known for Earthquakes and it is overdue for a major shaking. In 1989's Loma Prieta earthquake, the world saw what liquefaction can do, as blocks of the Marina District were reduced to rubble. Some properties are on top of fault lines like the ones damaged in South Napa that had fault traces running through them. When the ground actually splits during an earthquake, it can damage buildings and utility lines far beyond what shaking can do. Properties that are prone to "ground failure" are required by law to disclose that information to potential buyers which could potentially change buyer’s decision scenario.

Apart from earthquake hazard, 2018 wildfires have ravaged more than 1.2 million acres, destroyed more than 1200 homes in California. Housing market changed after that devastation in Northern California.

Properties that are prone to "ground failure" are required by law to disclose that information to potential buyers -- but the information isn't always easy to find until the transaction is already underway.

2. Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

Clients:

This study has been motivated by the need for a variety of stakeholders, including residential valuers, real estate agents, policy makers, insurance providers and mortgage lenders, to be able to accurately estimate the way in which changing perceptions of the risk of damage could alter property values.

Benefits to Clients:

(a) Real estate websites:

For example, if this study is made available to real estate companies such as Zillow or Trulia, then it would result in following business value additions:

- Zillow customers would be more informed on natural hazard attributes before buying a property.
- Zillow customers would use this hazard attributes as a key element, while making a decision on property purchase.
- Estimating the value of property by including various hazard attributes would yield higher precision of forecasting, which in turn leads to customer satisfaction and retention for the firm.
- Number of customers visiting Zillow would increase due to inclusion of hazard attributes compared to its competitors.
- Zillow makes money through Ad Sales as well. So providing natural hazard information for a property would invite advertisers from related companies such as seismic retrofit construction firms, earthquake insurance providers, etc to display their ads on the same page of those properties.
- Due to increase the volume of customers visiting the website, that result in increased number of premium real estate agents sign up for subscription services.
- With more customer engagement, cost per click from mortgage lender would get increased.

(b) Insurance Providers:

- With this information on natural hazard attribute tied up with property, insurance providers are less likely to encounter a loss on property claimed on those attributes. Because the insurance premium of the property would include this risk factor as well.

3. What data are you using? How will you acquire the data?

House Data:

For this project, Zillow property data for San Jose city will be collected and the method is outlined below:

- Zillow property data:
  Planning to use web scraping to get access the Zillow property data using python and web scraping packages such as selenium and BeautifulSoup.
  https://www.zillow.com/homes/san-jose_rb/

Natural Hazards Data:

In this project, natural hazards data such fault zone, landslide, liquefaction and fire hazard data will be collected from below sources:

- Earthquake (fault, liquefaction and landslide zones) hazard data:
  https://spatialservices.conservation.ca.gov/arcgis/rest/services/CGS_Earthquake_Hazard_Zones/SHP_ZoneInfo/MapServer
- Wildfire hazard potential data:
  http://www.fire.ca.gov/fire_prevention/fhsz_maps/FHSZ/santa_clara/San_Jose.pdf

Data Collection:

House properties data from Zillow:

- Zillow website API has inbuilt limitation on extracting data from Zillow. Therefore, web scraping will be done to extract the fields for single family homes and townhouse for “sold” and “for sale”(including By agent, By owner, New construction, Foreclosures, Coming Soon).
- Planning to extract property attributes such as latitude, longitude, address, zip, bedrooms, bathrooms, sqft, lot_size, year_built, price, sale_type, zestimate, date_sold, days_on_zillow, house_type and url using web scraping packages.

Earthquake Hazard data from CGS(California Geological Survey)

- CGS has recently developed web app for predicting earthquake hazard data for California. In this project, web app, sql, python requests and json packages will be used to extract earthquake hazard data from hazard web app.
- Zillow website API has inbuilt limitation on extracting data from Zillow. Therefore, web scraping will be done to extract the fields for single family homes and townhouse for “sold” and “for sale”(including By agent, By owner, New construction, Foreclosures, Coming Soon).

Fire Hazard Severity Zone data from California Fire Department

- Fire hazard data is available in pdf. San Jose city has very limited area comes under fire hazard severity zone. Parcel number and address will be extracted only for the area comes under fire hazard severity zone.

After collecting all property data and natural hazards data for San Jose, merge and append methods will be used to join data based on the address and parcel number attributes.

4. Briefly outline how you'll solve this problem.

Outline:

1. Data collection

   Tools: Python, selenium, BeautifulSoup, sql, python Requests, json packages, jupyter notebook

2. Data wrangling

   Tools: Python, jupyter notebook, Matplotlib, pandas, numpy

3. Exploratory analysis

   Tools: Python, jupyter notebook, Matplotlib, pandas, numpy

4. Visualization

   Tools: Python, matplotlib, folium, basemap and seaborn

5. Developing prediction models

   Tools: scikit

6. Summary and Report

   Tools: Jupyter notebook

Outcomes:

In this project, hazard and non-hazard properties in single family and townhouse will be analysed following segmentation approach.

- Geospatial visualization of natural hazard homes in single family and townhouse
- Most common features in single family (bedrooms, bathrooms,year_built, liquefaction, fault zone, landslide, fire hazard)
- Most common features in townhouse (bedrooms, bathrooms, year_built, liquefaction, fault zone, landslide, fire hazard)
- 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

5. What are your deliverables?

Deliverables include code, detailed report and slide deck.

DATA COLLECTION AND DATA WRANGLING

Introduction:

Initial challenge was getting data which was not readily available like in Kaggle website or some other websites for data extraction. After extracting data from various sources with different data wrangling techniques, next challenge was merging different data frames with python merge & append methods and formed into single analyzable data frame. In the following section, data collection and data wrangling methods will be discussed in detail.

Selection of City:

Before collecting data, first question came into my mind was which city would be appropriate for my analysis? The city was selected based on natural hazard level. Initially the plan was choosing San Francisco based population, area and hazard level. But, after going through map details of Earthquake hazard zone and fire hazard, San Jose city was selected instead of San Francisco(SFO). The reason behind for this selection were:

1. SFO (north) does not have underlying fault zone but SFO(south) has underlying fault zone. Also CGS did not evaluate SFO (south) for liquefaction and landslide. San Jose has underlying fault zones, liquefaction and landslide hazards (Fig 1). Pretty much San Jose covers all hazards.
2. Fire hazard was also another reason. After looking into fire hazard severity zone map for SFO, it was found that high fire hazard severity zone did come under SFO. But some part in San Jose East fell under high fire hazard severity zone.

3. Population and area wise san Jose city is similar to San Francisco.
4. San Jose Housing market is also much demandable like San Francisco market.

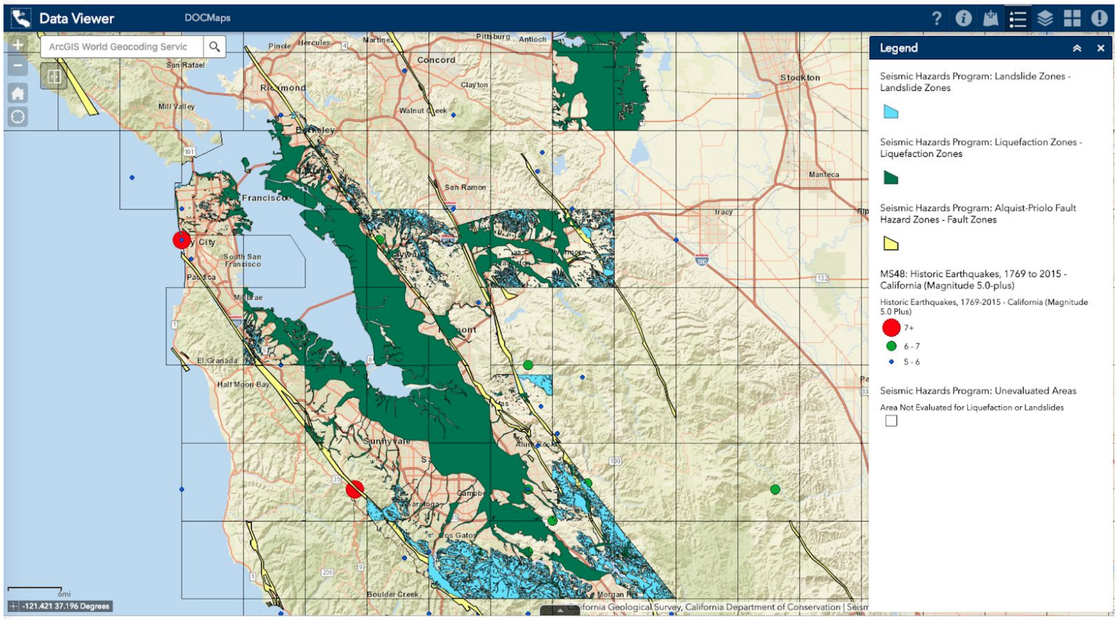

Data Viewer
DOCMaps
ArcGIS World Geocoding Servic
Legend
Seismic Hazards Program: Landslide Zones - Landslide Zones
Seismic Hazards Program: Liquefaction Zones - Liquefaction Zones
Seismic Hazards Program: Alquist-Priolo Fault Hazard Zones - Fault Zones
MS48: Historic Earthquakes, 1769 to 2015 - California (Magnitude 5.0-plus)
Historic Earthquakes, 1769-2015 - California (Magnitude 5.0 Plus)
7+
6 - 7
5 - 6
Seismic Hazards Program: Unevaluated Areas
Area Not Evaluated for Liquefaction or Landslides
San Rafael
Mill Valley
Richmond
Berkeley
Concord
Clayton
Walnut Creek
Pittsburg
Antioch
Stockton
Manteca
Tracy
San Ramon
Francisco
South San Francisco
Pacifica
Millbrae
El Granada
Half Moon Bay
Sunnyvale
Los Gatos
Boulder Creek
Morgan Hill
-121.421 37.196 Degrees
California Geological Survey, California Department of Conservation

Fig 1: Earthquake Hazard zones map

Data Source:

House Data:

For this project, Zillow property data for San Jose city were collected and the method is outlined below:

1. *Zillow property data:*

   Web scraping was used to get access the Zillow property data using python and web scraping packages such as selenium and BeautifulSoup.
   https://www.zillow.com/homes/san-jose_rb/

Natural Hazards Data:

In this project, natural hazards data such fault zone, landslide, liquefaction and fire hazard data were collected from below sources:

2. *Earthquake (fault, liquefaction and landslide zones) hazard data:*

   https://spatialservices.conservation.ca.gov/arcgis/rest/services/CGS_Earthquake_Hazard_Zones/SHP_ZoneInfo/MapServer

3. *Wildfire hazard potential data:*

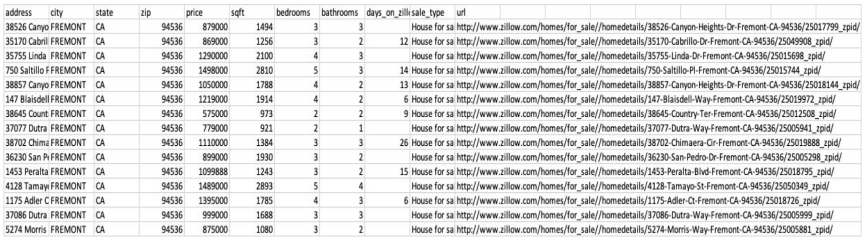
http://www.fire.ca.gov/fire_prevention/fhsz_maps/FHSZ/santa_clara/San_Jose.pdf

Data Collection:

House properties data from Zillow:

Zillow website API has inbuilt limitation on extracting data from Zillow. Also, Zillow API does not have provision to download zip code-wise/city-wise data. Therefore, it was decided to do web scraping the following fields for single family properties and townhouse for sold and for sale(By agent,By owner, New construction, Foreclosures, Coming Soon) for different zip codes in San Jose. Latitude, longitude, address, zip, bedrooms, bathrooms, sqft, lot_size, year_built, price, sale_type, zestimate, date_sold, days_on_zillow, house_type, url were extracted.

Python code which imports Selenium and BeautifulSoup packages to scrape my Zillow features for all zip codes in San Jose was written. Below is the sample scraped data from Zillow (Fig 2)

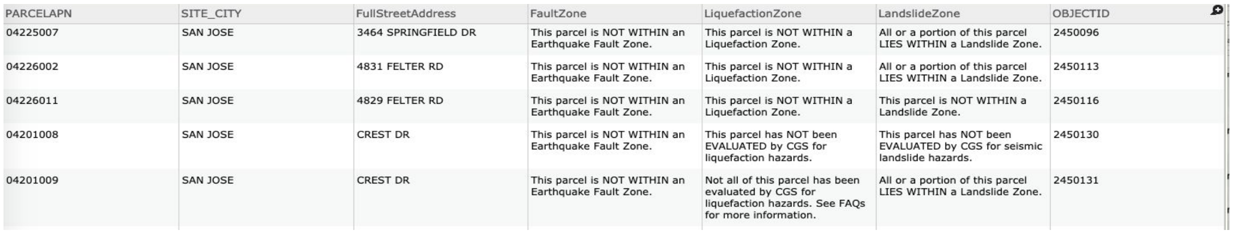


address	city	state	zip	price	sqft	bedrooms	bathrooms	days_on_zill	sale_type	url
38526 Canyo	FREMONT	CA	94536	879000	1494	3	3		House for sa	http://www.zillow.com/homes/for_sale//homedetails/38526-Canyon-Heights-Dr-Fremont-CA-94536/25017799_zpid/
35170 Cabril	FREMONT	CA	94536	869000	1256	3	2	12	House for sa	http://www.zillow.com/homes/for_sale//homedetails/35170-Cabrillo-Dr-Fremont-CA-94536/25049908_zpid/
35755 Linda	FREMONT	CA	94536	1290000	2100	4	3		House for sa	http://www.zillow.com/homes/for_sale//homedetails/35755-Linda-Dr-Fremont-CA-94536/25015698_zpid/
750 Saltillo F	FREMONT	CA	94536	1498000	2810	5	3	14	House for sa	http://www.zillow.com/homes/for_sale//homedetails/750-Saltillo-Pl-Fremont-CA-94536/25015744_zpid/
38857 Canyo	FREMONT	CA	94536	1050000	1788	4	2	13	House for sa	http://www.zillow.com/homes/for_sale//homedetails/38857-Canyon-Heights-Dr-Fremont-CA-94536/25018144_zpid/
147 Blaisdell	FREMONT	CA	94536	1219000	1914	4	2	6	House for sa	http://www.zillow.com/homes/for_sale//homedetails/147-Blaisdell-Way-Fremont-CA-94536/25019972_zpid/
38645 Count	FREMONT	CA	94536	575000	973	2	2	9	House for sa	http://www.zillow.com/homes/for_sale//homedetails/38645-Country-Ter-Fremont-CA-94536/25012508_zpid/
37077 Dutra	FREMONT	CA	94536	779000	921	2	1		House for sa	http://www.zillow.com/homes/for_sale//homedetails/37077-Dutra-Way-Fremont-CA-94536/25005941_zpid/
38702 Chima	FREMONT	CA	94536	1110000	1384	3	3	26	House for sa	http://www.zillow.com/homes/for_sale//homedetails/38702-Chimaera-Cir-Fremont-CA-94536/25019888_zpid/
36230 San P	FREMONT	CA	94536	899000	1930	3	2		House for sa	http://www.zillow.com/homes/for_sale//homedetails/36230-San-Pedro-Dr-Fremont-CA-94536/25005298_zpid/
1453 Peralta	FREMONT	CA	94536	1099888	1243	3	2	15	House for sa	http://www.zillow.com/homes/for_sale//homedetails/1453-Peralta-Blvd-Fremont-CA-94536/25018795_zpid/
4128 Tamay	FREMONT	CA	94536	1489000	2893	5	4		House for sa	http://www.zillow.com/homes/for_sale//homedetails/4128-Tamayo-St-Fremont-CA-94536/25050349_zpid/
1175 Adler C	FREMONT	CA	94536	1395000	1785	4	3	6	House for sa	http://www.zillow.com/homes/for_sale//homedetails/1175-Adler-Ct-Fremont-CA-94536/25018726_zpid/
37086 Dutra	FREMONT	CA	94536	999000	1688	3	3		House for sa	http://www.zillow.com/homes/for_sale//homedetails/37086-Dutra-Way-Fremont-CA-94536/25005999_zpid/
5274 Morris	FREMONT	CA	94536	875000	1080	3	2		House for sa	http://www.zillow.com/homes/for_sale//homedetails/5274-Morris-Way-Fremont-CA-94536/25005881_zpid/

Fig 2: Sample scraped data from Zillow

Earthquake hazard Zone Data:

Sql code was used to extract data from California geological survey (CGS) web app application for seismic hazard data (Fig 2). But the problem was, CGS limited to retrieve only 1000 rows at a time with all attributes using sql query. It was decided to fetch object ids only using python program, then using those object IDs fetched the actual rows.



PARCELAPN	SITE_CITY	FullStreetAddress	FaultZone	LiquefactionZone	LandslideZone	OBJECTID
04225007	SAN JOSE	3464 SPRINGFIELD DR	This parcel is NOT WITHIN an Earthquake Fault Zone.	This parcel is NOT WITHIN a Liquefaction Zone.	All or a portion of this parcel LIES WITHIN a Landslide Zone.	2450096
04226002	SAN JOSE	4831 FELTER RD	This parcel is NOT WITHIN an Earthquake Fault Zone.	This parcel is NOT WITHIN a Liquefaction Zone.	All or a portion of this parcel LIES WITHIN a Landslide Zone.	2450113
04226011	SAN JOSE	4829 FELTER RD	This parcel is NOT WITHIN an Earthquake Fault Zone.	This parcel is NOT WITHIN a Liquefaction Zone.	This parcel is NOT WITHIN a Landslide Zone.	2450116
04201008	SAN JOSE	CREST DR	This parcel is NOT WITHIN an Earthquake Fault Zone.	This parcel has NOT been EVALUATED by CGS for liquefaction hazards.	This parcel has NOT been EVALUATED by CGS for seismic landslide hazards.	2450130
04201009	SAN JOSE	CREST DR	This parcel is NOT WITHIN an Earthquake Fault Zone.	Not all of this parcel has been evaluated by CGS for liquefaction hazards. See FAQs for more information.	All or a portion of this parcel LIES WITHIN a Landslide Zone.	2450131

Fig 3: Sample seismic hazard data from CGS web application

Using python requests and json packages,All extracted data were stored in single csv file.
Total Features and attributes in dataframe: 261195 & 7

Fault zone, liquefaction Zone and landslide zone column string values were converted to 0,1,NA based on following condition simply for simplicity:

If Liquefaction Zone:

```
LIES WITHIN a Liquefaction Zone = 1
NOT been EVALUATED by CGS for liquefaction hazards = NA
NOT within a liquefaction zone = 0
```

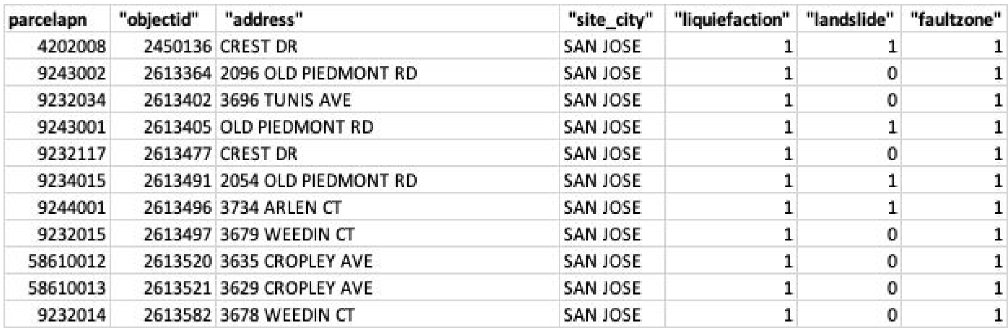
If landslide Zone:

```
LIES WITHIN a Landslide Zone = 1
NOT been EVALUATED by CGS for seismic landslide hazard = NA
NOT within a landslide zone = 0
```

If Fault Zone:

```
LIES WITHIN an Earthquake Fault Zone = 1
NOT WITHIN an Earthquake Fault Zone = 0
Not been EVALUATED by CGS  = NA
```

After converting features as explained above, sample seismic hazard data was stored in csv format and it is shown in Fig 4.



parcelapn	"objectid"	"address"	"site_city"	"liquiefaction"	"landslide"	"faultzone"
4202008	2450136	CREST DR	SAN JOSE	1	1	1
9243002	2613364	2096 OLD PIEDMONT RD	SAN JOSE	1	0	1
9232034	2613402	3696 TUNIS AVE	SAN JOSE	1	0	1
9243001	2613405	OLD PIEDMONT RD	SAN JOSE	1	1	1
9232117	2613477	CREST DR	SAN JOSE	1	0	1
9234015	2613491	2054 OLD PIEDMONT RD	SAN JOSE	1	1	1
9244001	2613496	3734 ARLEN CT	SAN JOSE	1	1	1
9232015	2613497	3679 WEEDIN CT	SAN JOSE	1	0	1
58610012	2613520	3635 CROPLEY AVE	SAN JOSE	1	0	1
58610013	2613521	3629 CROPLEY AVE	SAN JOSE	1	0	1
9232014	2613582	3678 WEEDIN CT	SAN JOSE	1	0	1

Fig 4: Sample seismic hazard data

Fire hazard severity zone (FHSZ):

Fire hazard data is available in pdf . San Jose city has very limited area comes under fire hazard severity zone. Manually parcel numbers and address were extracted only for the area comes under fire hazard severity zone. Total features and attributes: 53 & 6. The fire hazard severity zone map can be obtained from http://www.fire.ca.gov/fire_prevention/fhsz_maps/FHSZ/santa_clara/San_Jose.pdf and it is shown in Fig 5a. Sample output data is shown in Fig 5b.

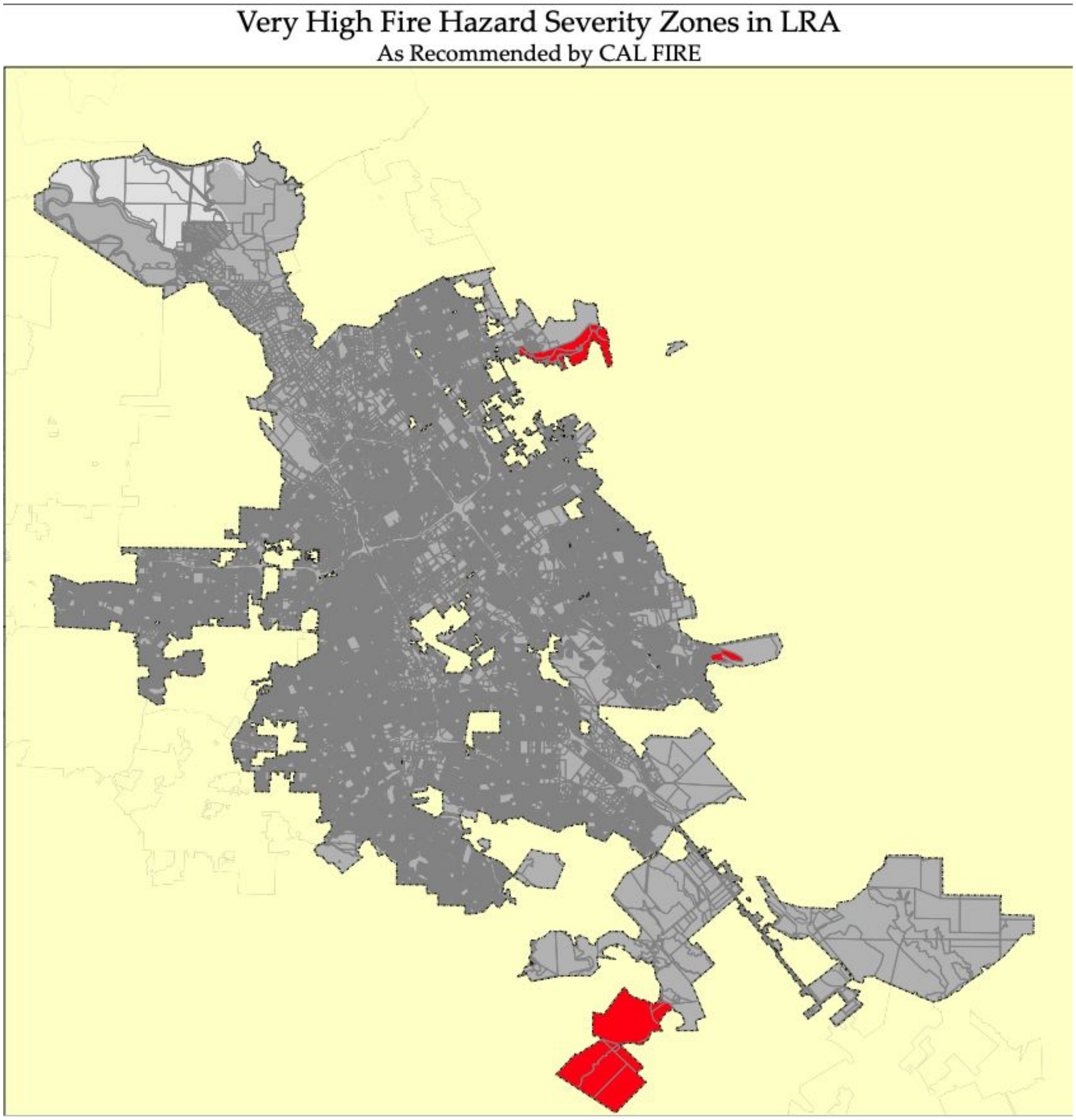

Very High Fire Hazard Severity Zones in LRA
As Recommended by CAL FIRE

Fig 5a: Very High Fire Hazard Severity Zones

Address	city	state	zip	parcel number	FHSZ in LRA
87 La Quinta Dr	San Jose	CA	95127	61254059	1
89 La Quinta Dr	San Jose	CA	95127	61254053	1
91 La Quinta Dr	San Jose	CA	95127	61254054	1
93 La Quinta Dr	San Jose	CA	95127	61254055	1
95 La Quinta Dr	San Jose	CA	95127	61254056	1
97 La Quinta Dr	San Jose	CA	95127	61254057	1
101 Spyglass Hill Rd	San Jose	CA	95127	61254001	1

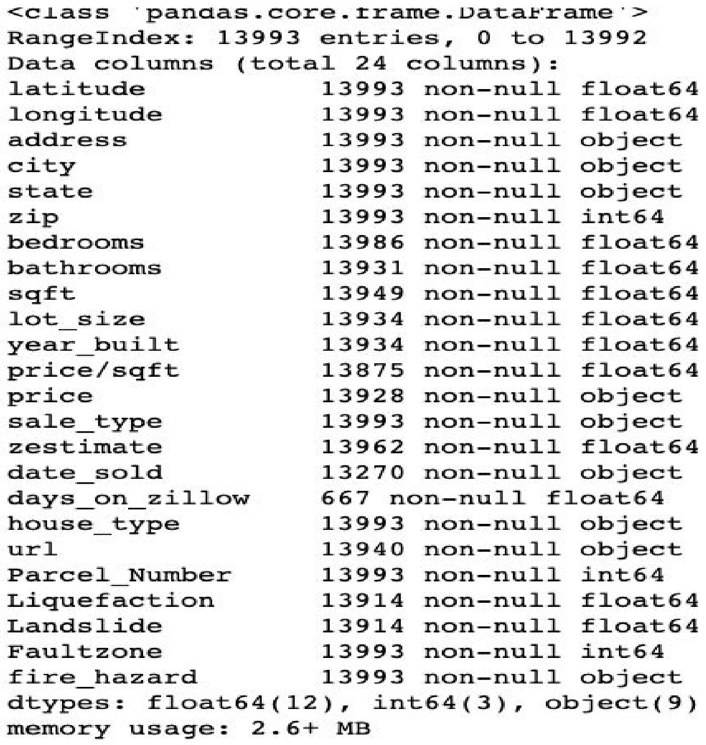
Fig 5b: Sample fire hazard data

Data Wrangling:

Merging Dataframes:

After reading Zillow properties, seismic hazard and fire hazard csv files saved in different dataframes, the next challenge was merging all data frames. The following steps were used for merging:

- ➢ Zillow properties and seismic hazard data based on 'address' column were merged. Then 'fire hazard' column in the merged data frame was added.
- ➢ Fire hazard and seismic hazard data were merged based on ‘address’ column.
- ➢ Above first merged data frame was appended to the second merged dataframe.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13993 entries, 0 to 13992
Data columns (total 24 columns):
latitude          13993 non-null float64
longitude         13993 non-null float64
address           13993 non-null object
city              13993 non-null object
state             13993 non-null object
zip               13993 non-null int64
bedrooms          13986 non-null float64
bathrooms         13931 non-null float64
sqft              13949 non-null float64
lot_size          13934 non-null float64
year_built        13934 non-null float64
price/sqft        13875 non-null float64
price             13928 non-null object
sale_type         13993 non-null object
zestimate         13962 non-null float64
date_sold         13270 non-null object
days_on_zillow    667 non-null float64
house_type        13993 non-null object
url               13940 non-null object
Parcel_Number     13993 non-null int64
Liquefaction      13914 non-null float64
Landslide         13914 non-null float64
Faultzone         13993 non-null int64
fire_hazard       13993 non-null object
dtypes: float64(12), int64(3), object(9)
memory usage: 2.6+ MB
```

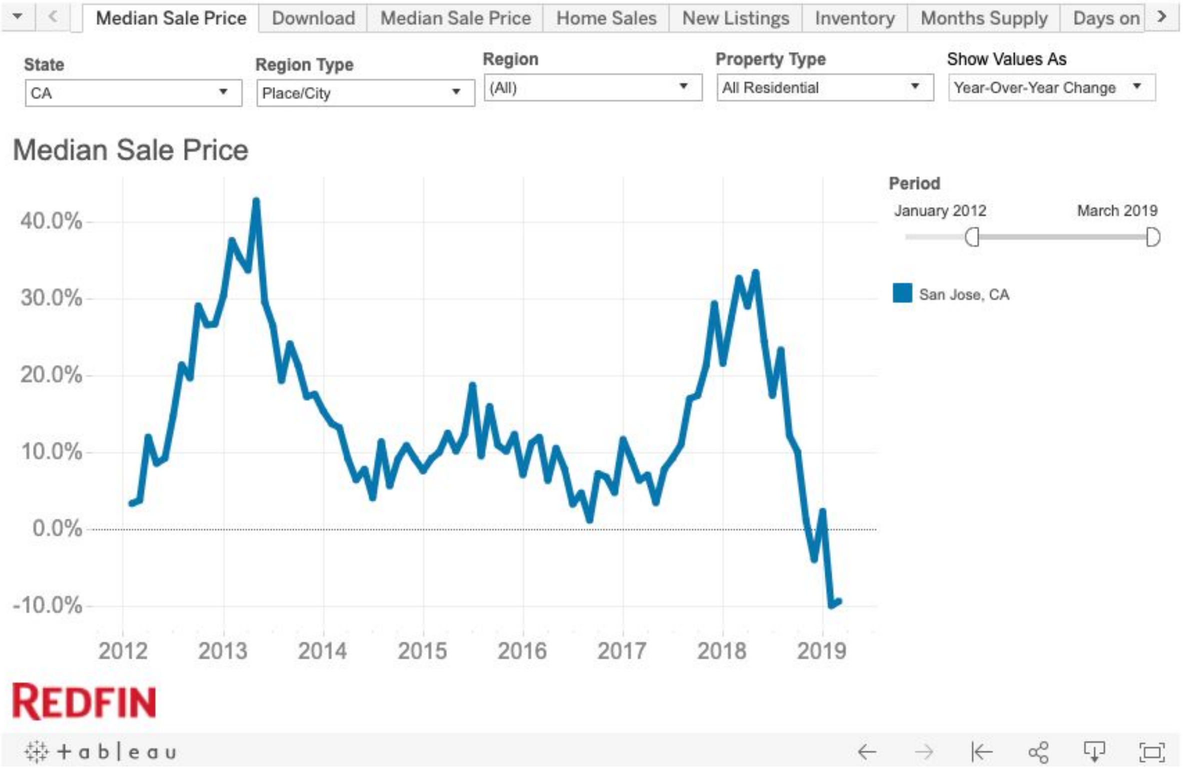
Fig 6: Merged dataframe info

Handling Duplicates:

- duplicates based on parcel number and dropped duplicates
- duplicates based on address and dropped duplicates
- out of 13993 rows 13617 were obtained after removing duplicates

Correcting Formats:

- It was noticed price column type was object instead of integer. The price column were reported as SOLD: $ --M. Strip method was used to remove text and converted unit M to dollar.
- It was also observed that few houses named as APT,UNT,# in the address column. But listed as single family home. The house type was changed to townhouse for those mentioned as APT,UNT,# in address column.
- Sold price were from 2016 to 2019. In order to normalize sold price, sold price was adjusted to current price based on redfin median sale price change over the period of time (Fig 7)



State: CA
Region Type: Place/City
Region: (All)
Property Type: All Residential
Show Values As: Year-Over-Year Change

Fig 7: Median Sale price change over the period of time (San Jose)

Correcting error in data:

- Few properties were observed with very low and very high sold price values. It was also cross checked with other websites such as redfin and trulia and replaced those values. 36 properties were replaced with correct sold price and sold date.
- Few properties were less than 100,000 and those properties were sold under non arm transactions type. These are not real property price. Hence, it was decided to remove those data by finding percentage difference between zestimate and adjusted sold price. It was decided to remove properties with price diff percentage of more than 38% based on frequency distribution plot.

Correcting data type:

- Date sold format was converted from object type to date time.

Handling Missing data:

- Year_built columns were replaced with median.
- Bedrooms with zero count were removed. Bedrooms greater than or equal to 1 were considered for this project. Studio rooms were not considered.
- Properties without bedrooms, bathrooms and sqft details were removed.
- Missing sqft properties were dropped.
- For missing bedrooms and bathrooms, ffill method was used based sorted sqft values.
- Missing lot size were filled with median values.
- Date sold were split into month and year for further exploratory analysis.

- All missing values were handled except zestimate and days on zillow.

Handling Outliers:

- Normalized price box plot (Fig 8) was plotted to see outliers in data. It was found one house more than 30 M and checked in Zillow. It was wrong data. It was removed from data.
- Similarly, few houses were more than 4 M as shown in the figure below. It was also cross checked with Zillow and found one house was not single family residential. It was removed from the data.
- Following procedures were followed to remove outliers:
  - The interquartile range for the data was calculated.
  - The interquartile range (IQR) was multiplied by the number 1.5.
  - 1.5 x (IQR) to the third quartile was added. Any number greater than this was a suspected outlier and removed.
  - 1.5 x (IQR) was subtracted from the first quartile. Any number greater than this was a suspected outlier and removed.

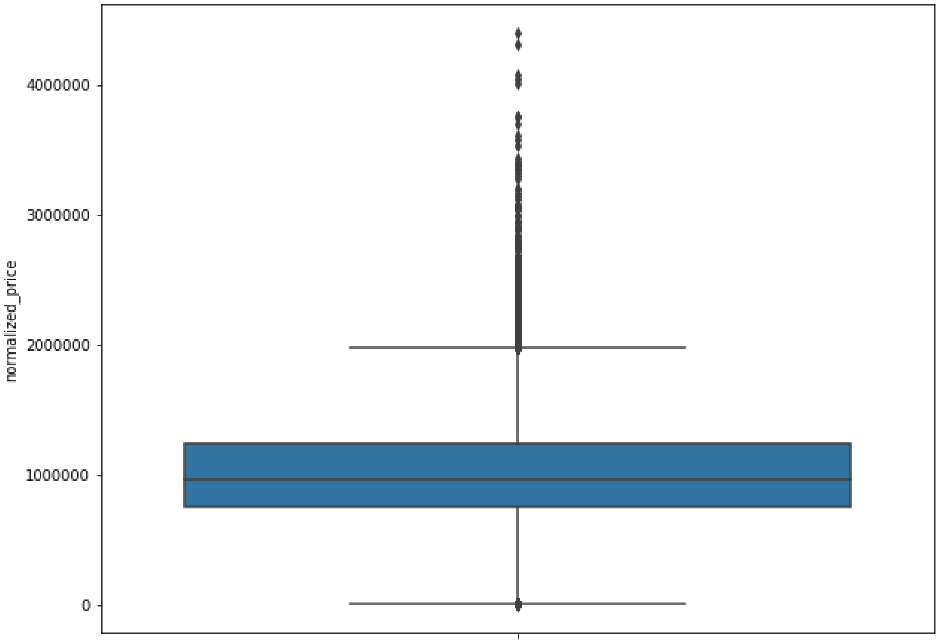


Fig 8: Box plot (adjusted price)

Jupyter notebook data collection notebook:
https://nbviewer.jupyter.org/github/umaraju18/CapStone-Project/blob/master/sanjose_data_collection.ipynb

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis:

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.

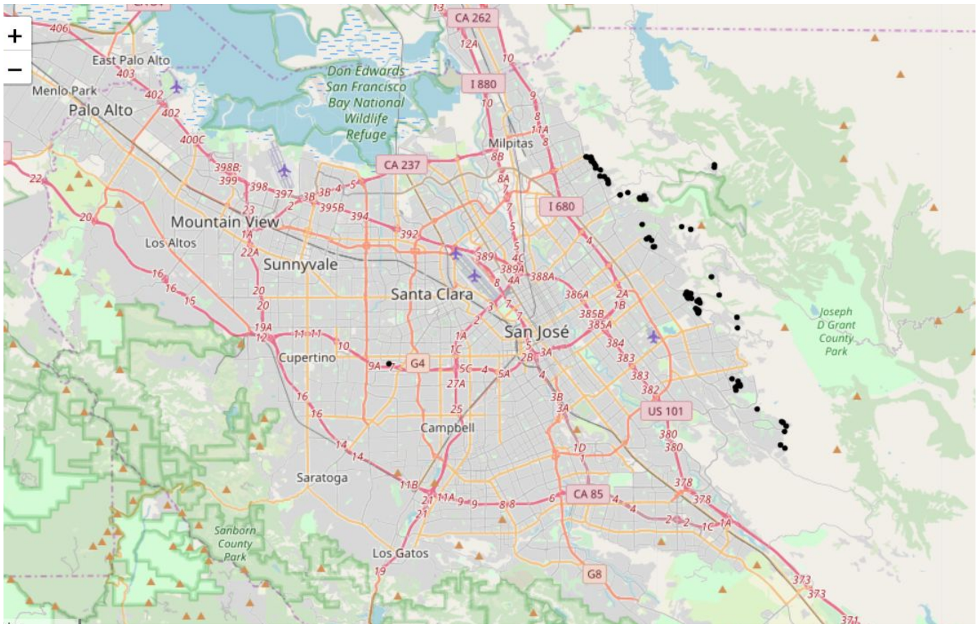


Fig 1a: Fault zone distribution (single family homes)

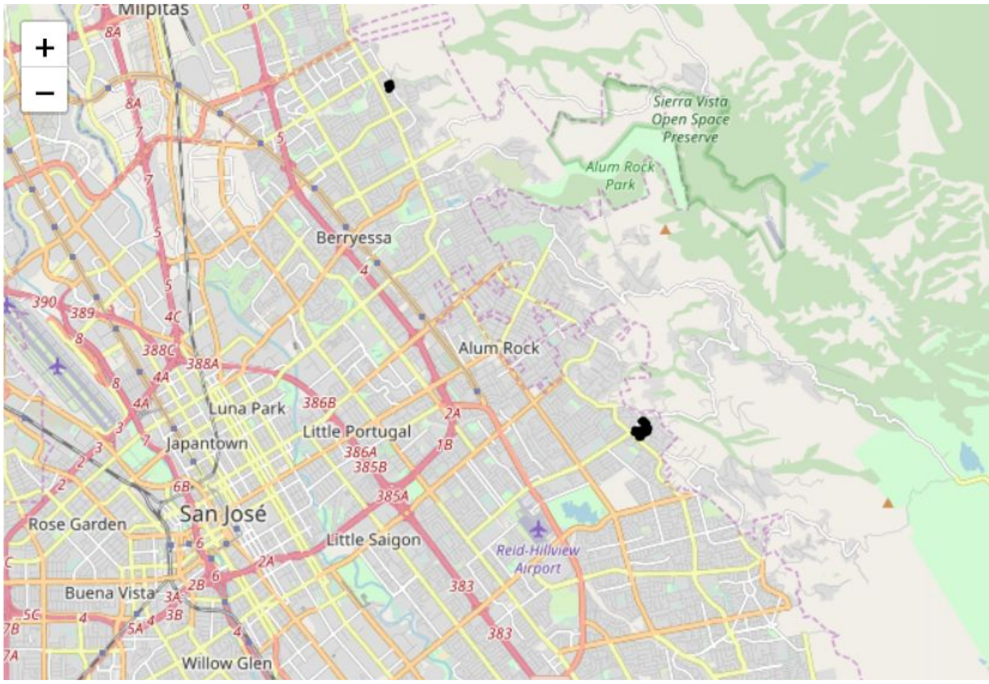


Fig 1b: Fault zone distribution (town homes)

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

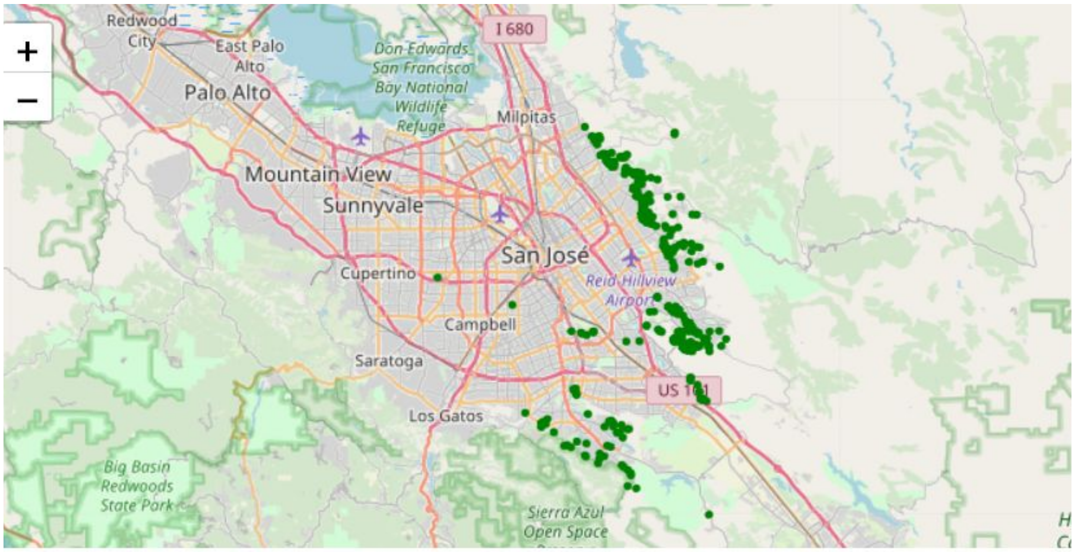


Fig 2a: Landslide distribution (single family homes)

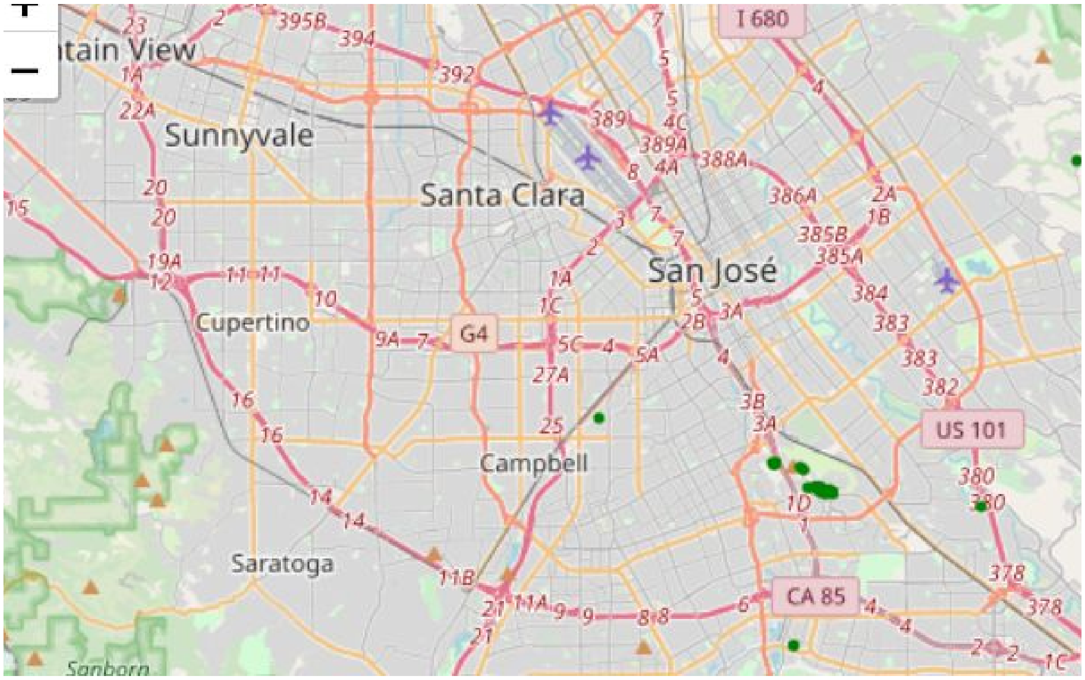


Fig 2b: Landslide distribution (town homes)

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

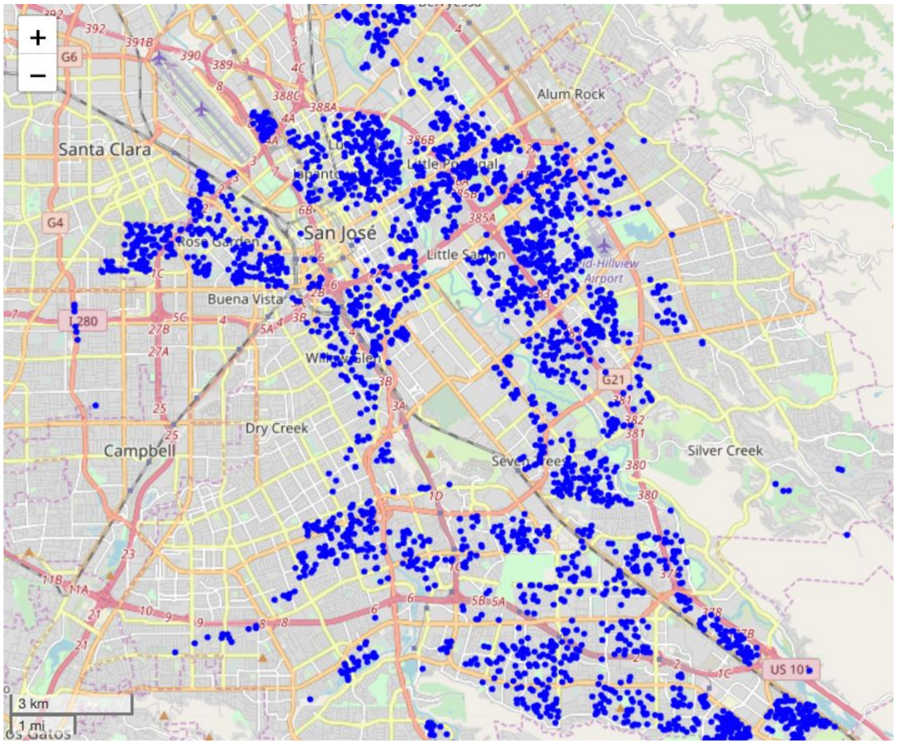


Fig 3a: Liquefaction distribution (single family homes)

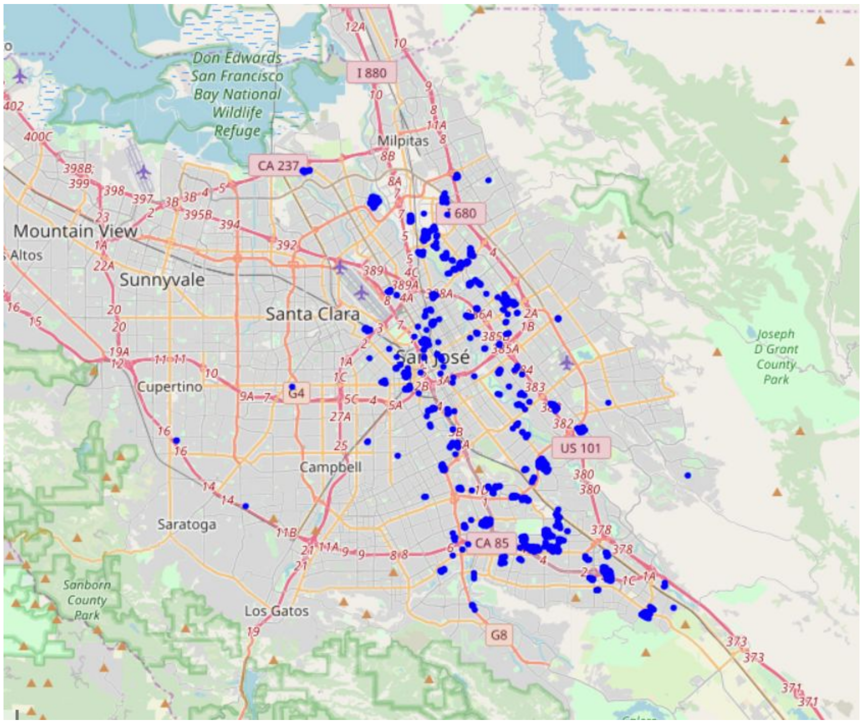


Fig 3b: Liquefaction distribution (town homes)

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

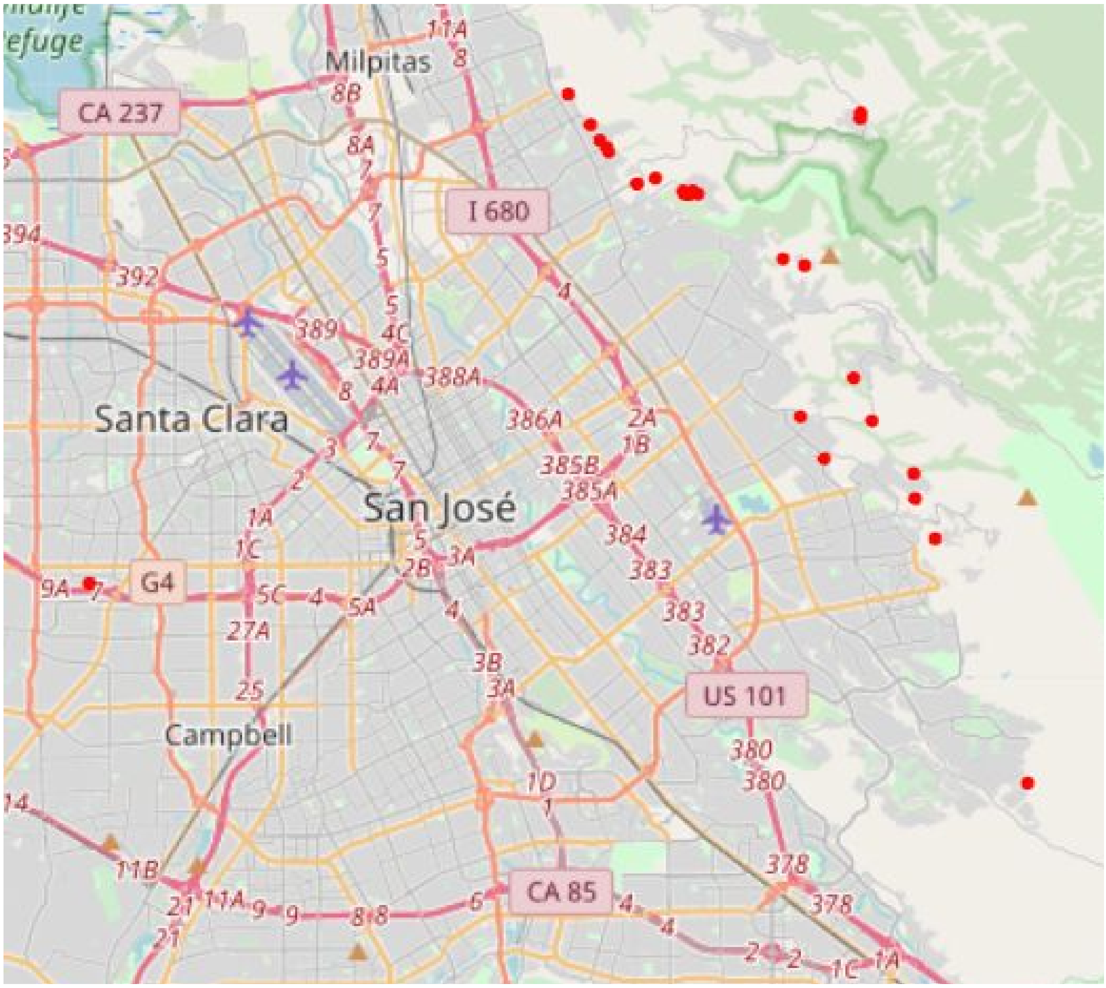


Fig 4: Liquefaction and landslide distribution (single family)

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?

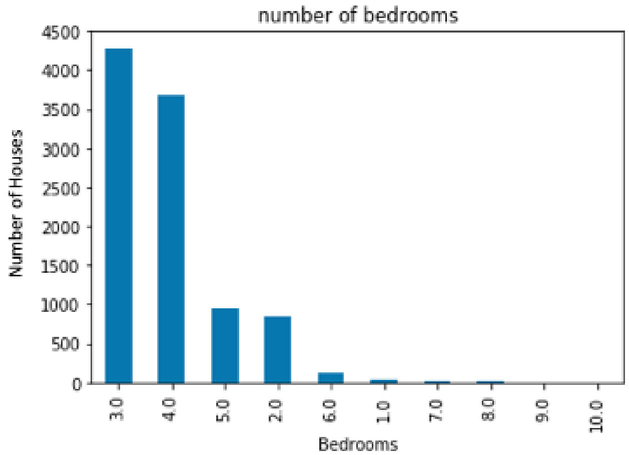


Fig 5a: Bedrooms vs count (single family)

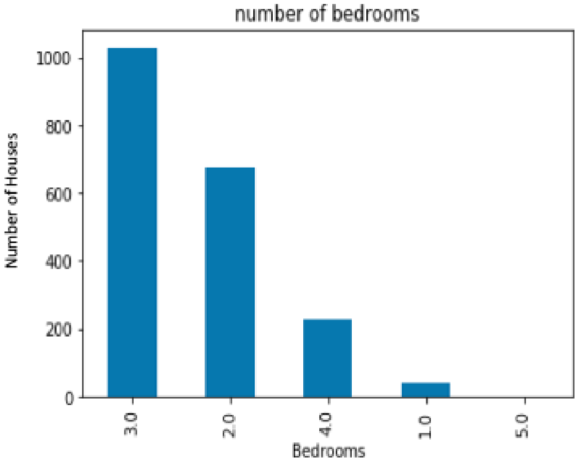


Fig 5b: Bedrooms vs count (townhomes)

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?

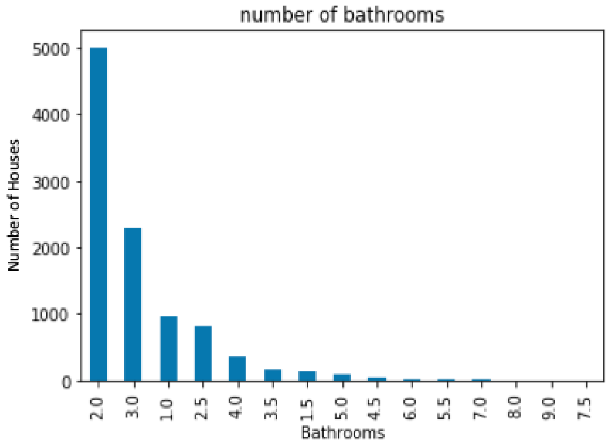


Fig 6a: Bathrooms vs count (single family)

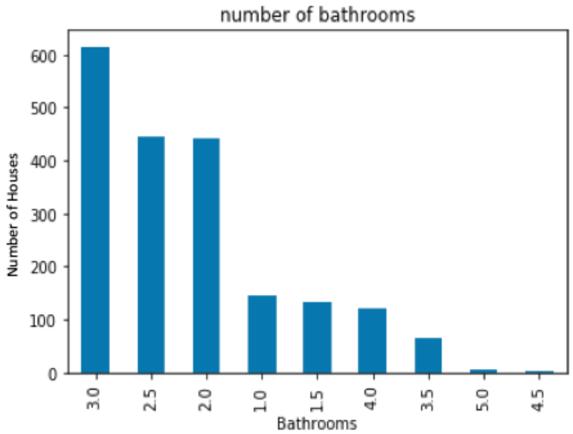


Fig 6b: Bathrooms vs count (townhomes)

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?

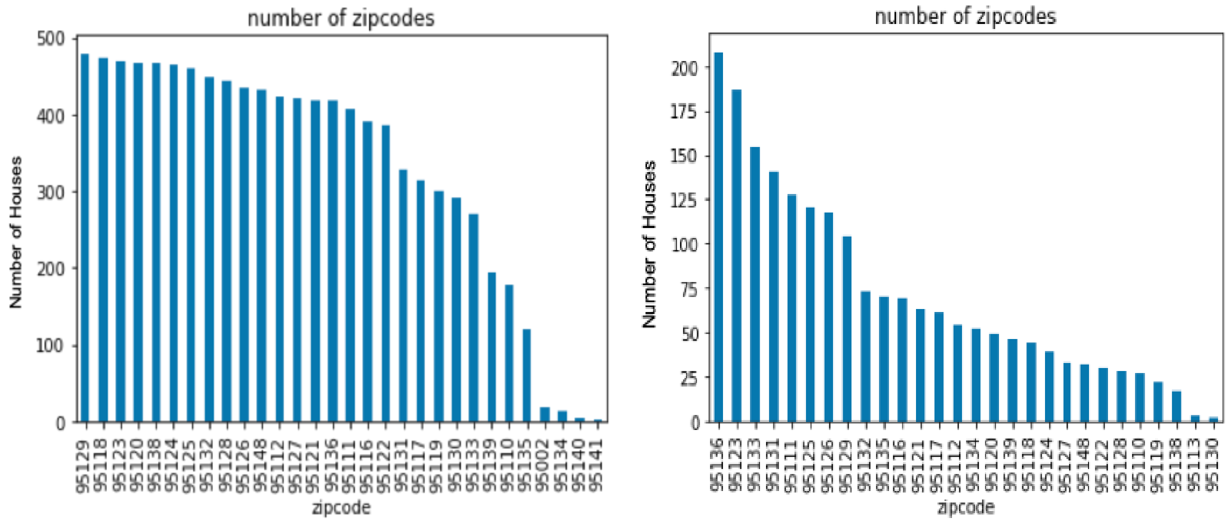


Fig 7a: Zip Code vs Number of Houses (single family) Fig 7b: Zip Code vs Number of Houses (townhomes)

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 ?

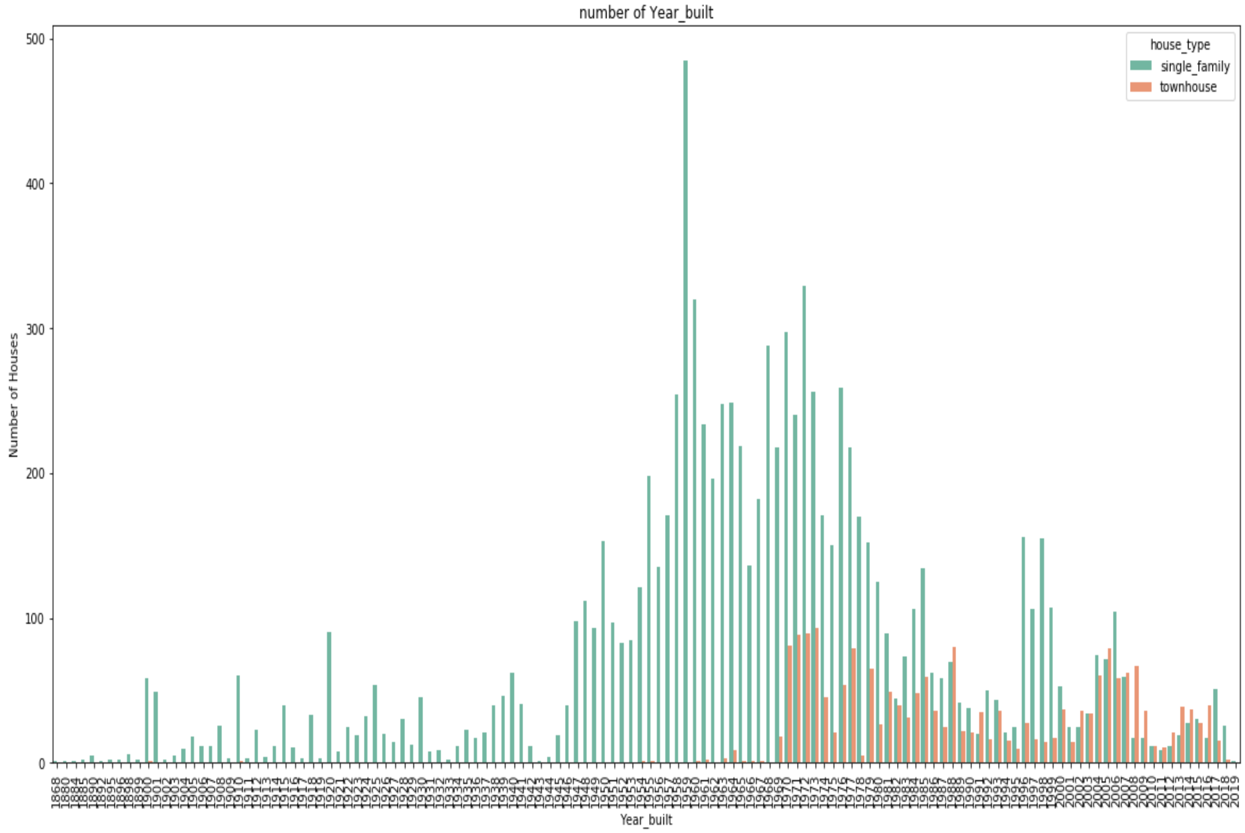


Fig 8: Year built vs Number of Houses (Single and townhomes)

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?

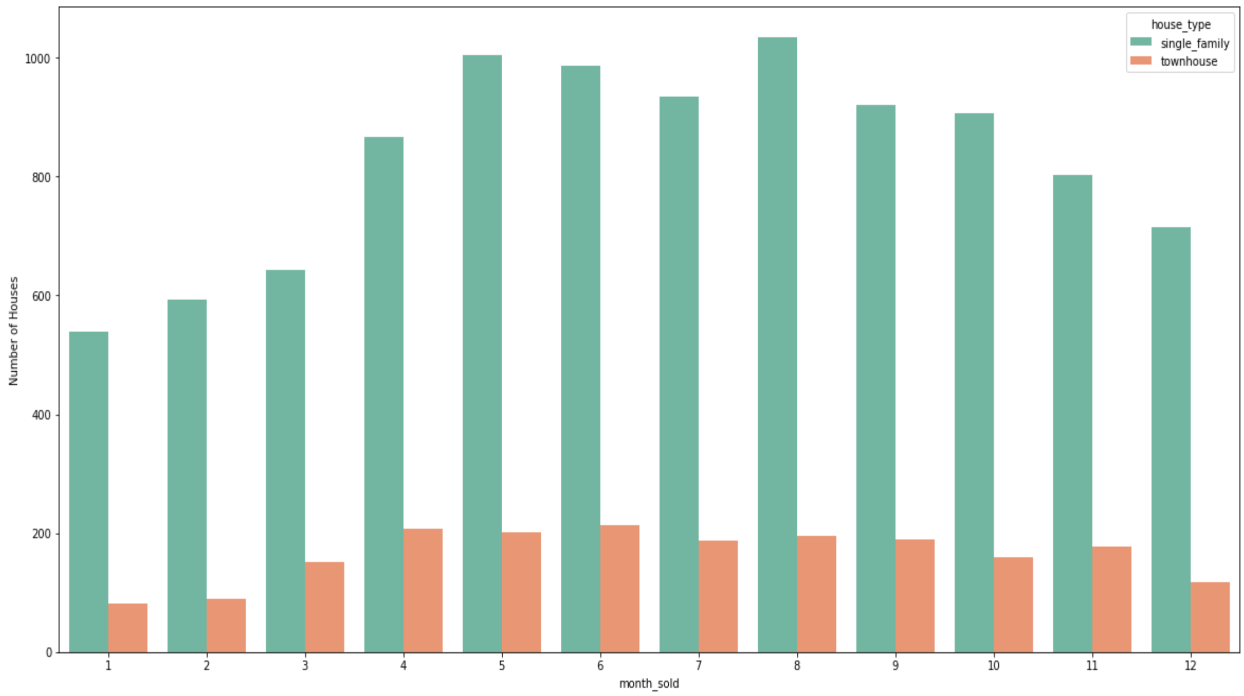


Fig 9: Month_sold vs Number of Houses (single family & townhouse)

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 sold?

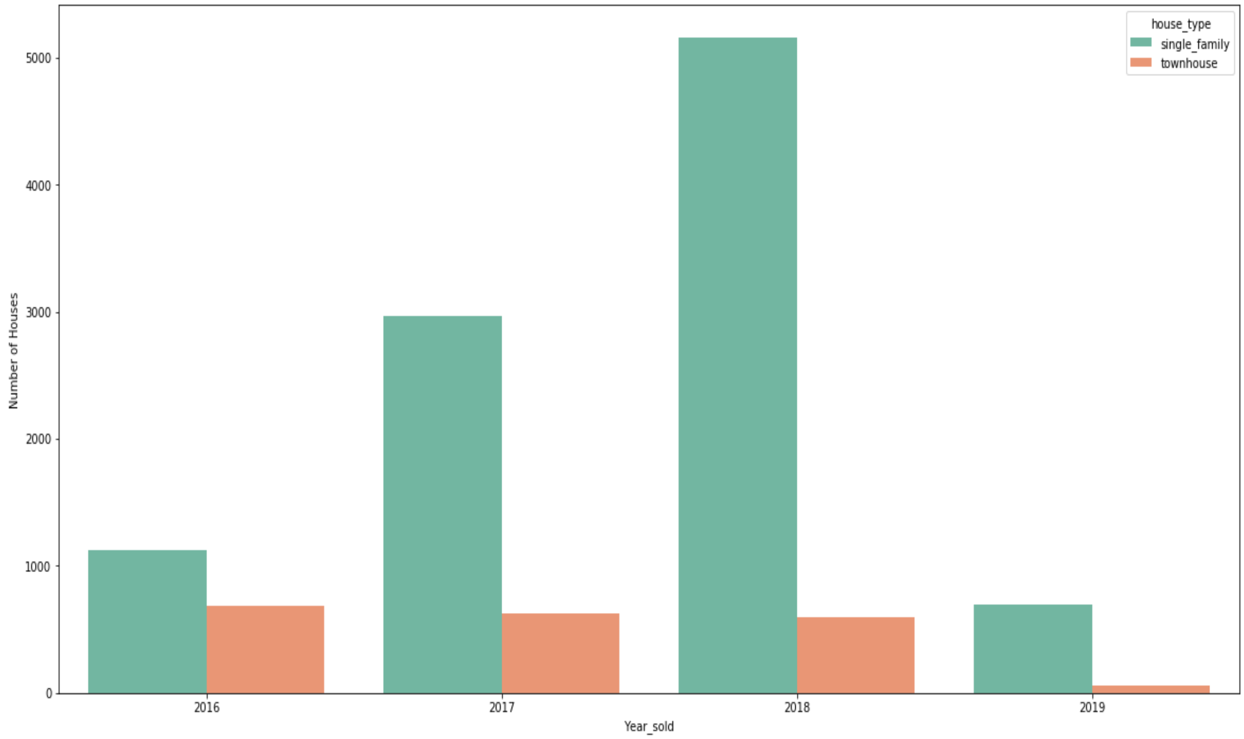


Fig 10: Year_sold vs Number of Houses(single family & Townhouse)

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?

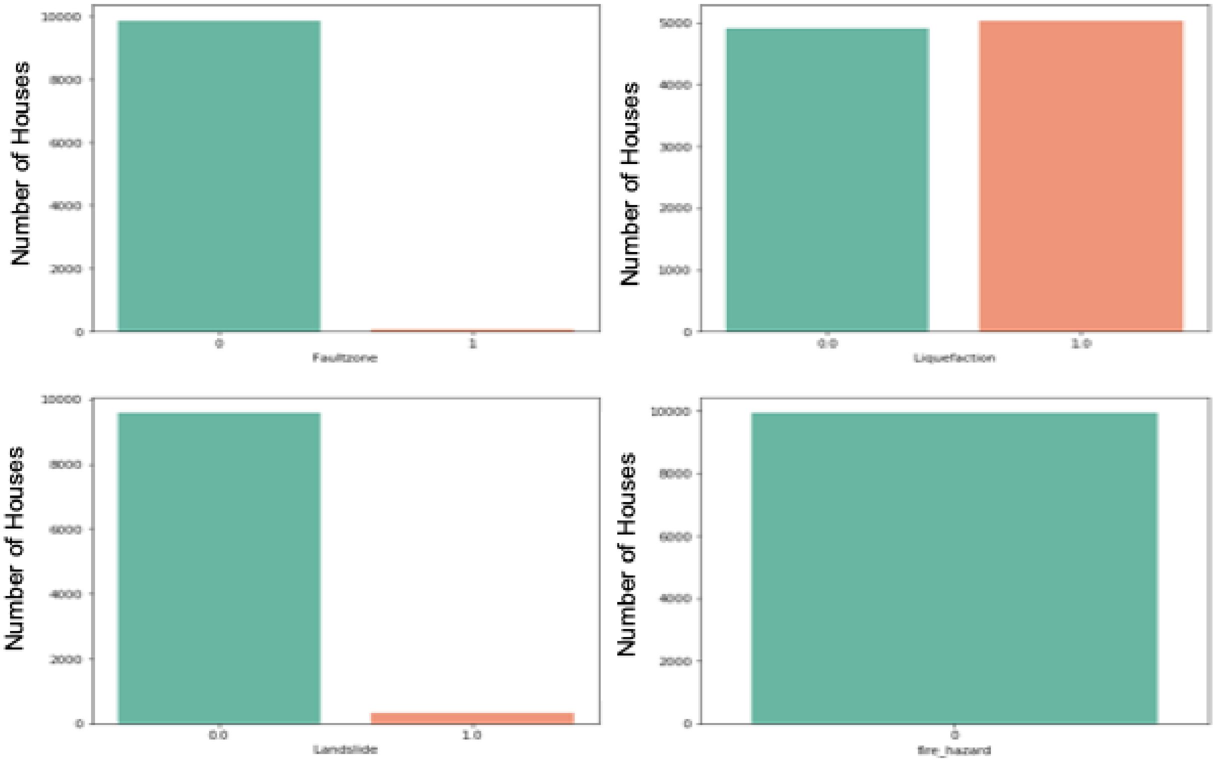


Fig 11a: Hazard vs Number of Houses (single family)

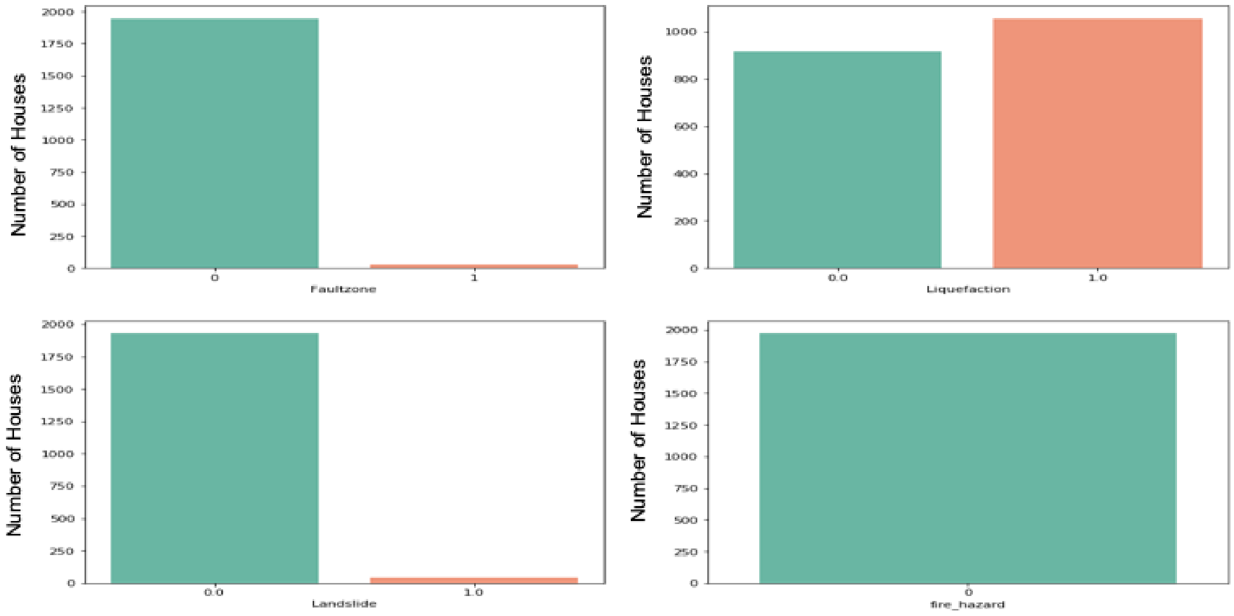


Fig 11b: Hazard vs Number of Houses (townhomes)

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?

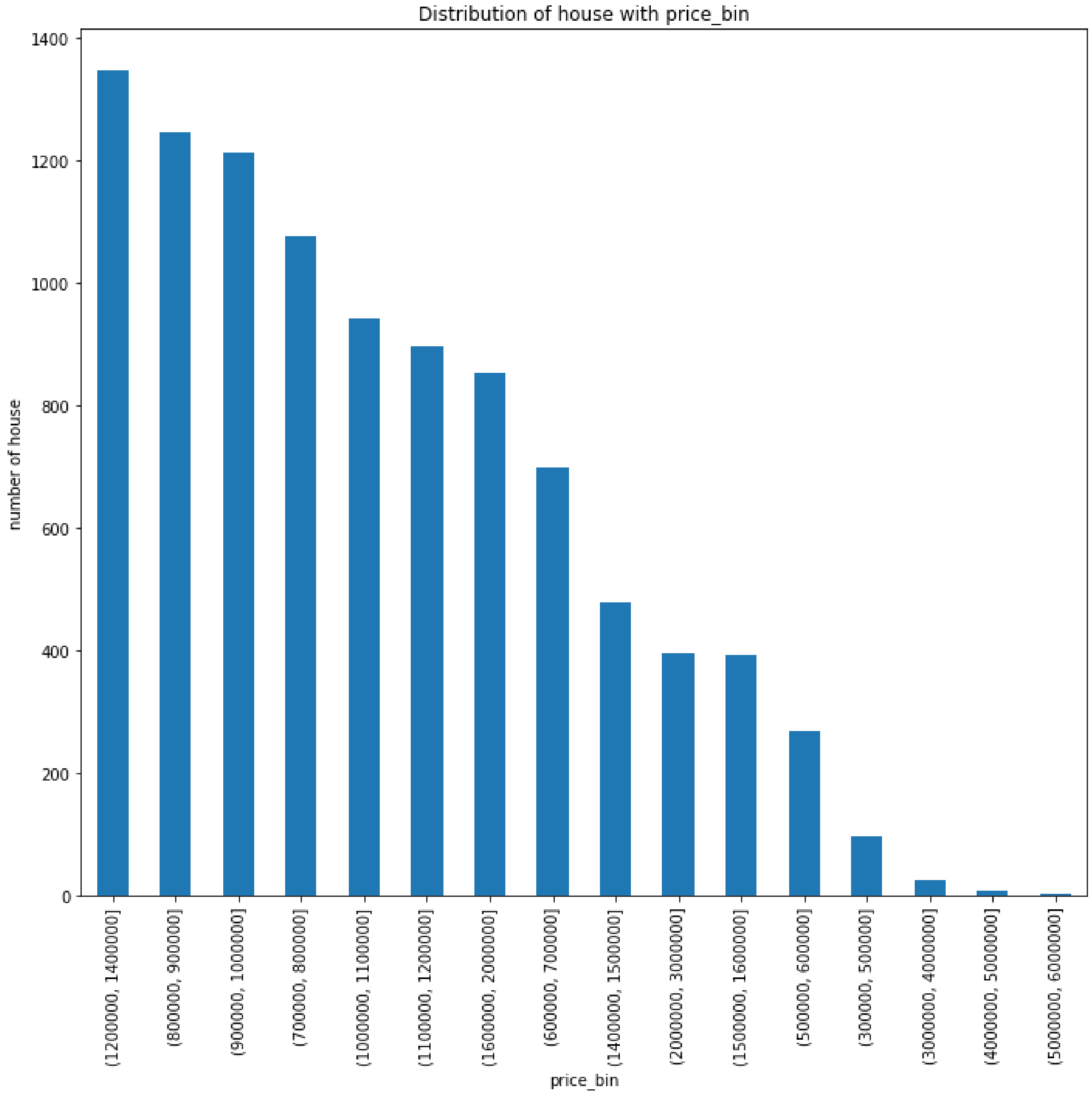


Fig 12a: Price_bin vs count (single family)

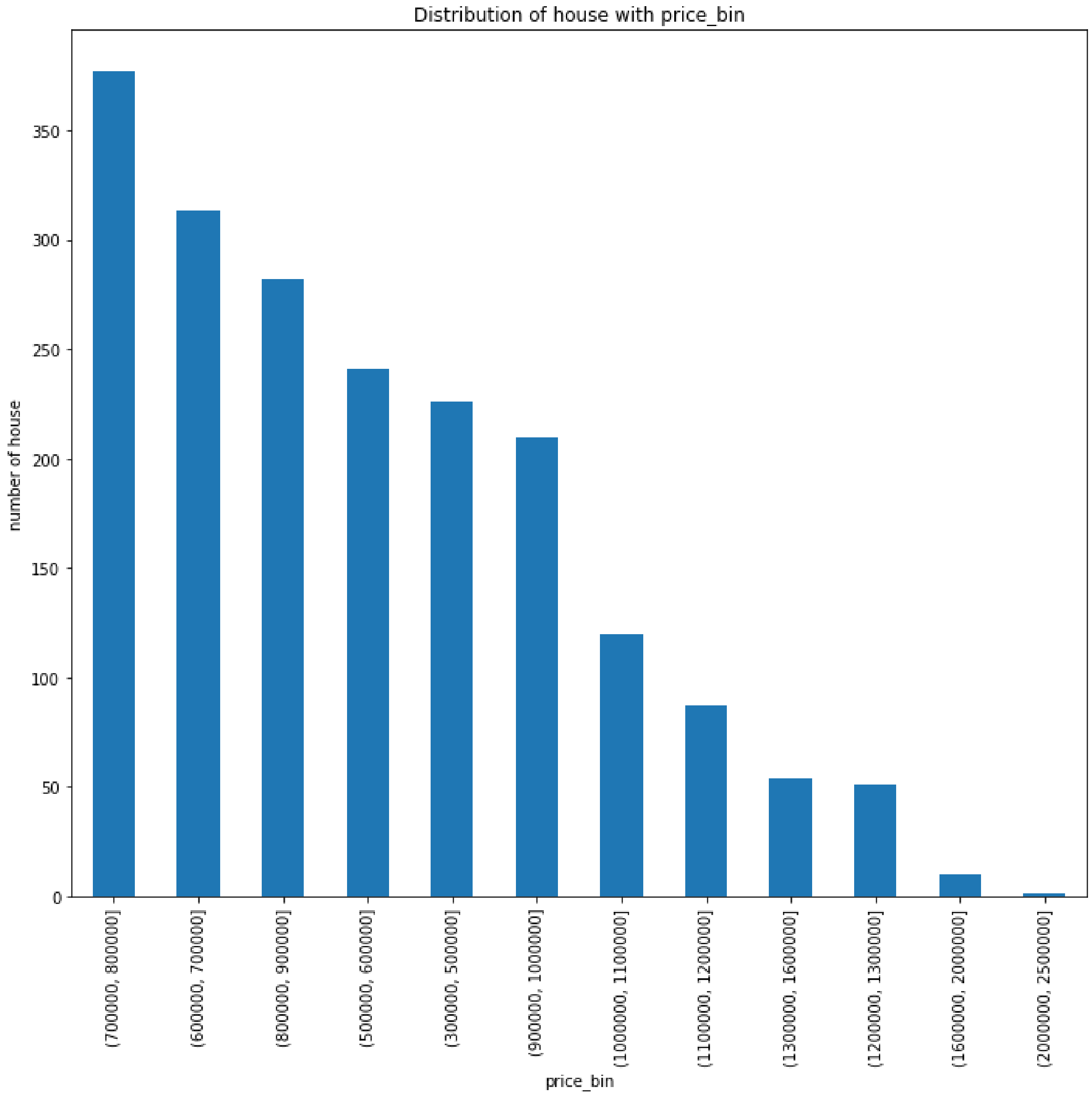


Fig 12b: Price_bin vs count (townhomes)

How many number of bedrooms in different price bin?

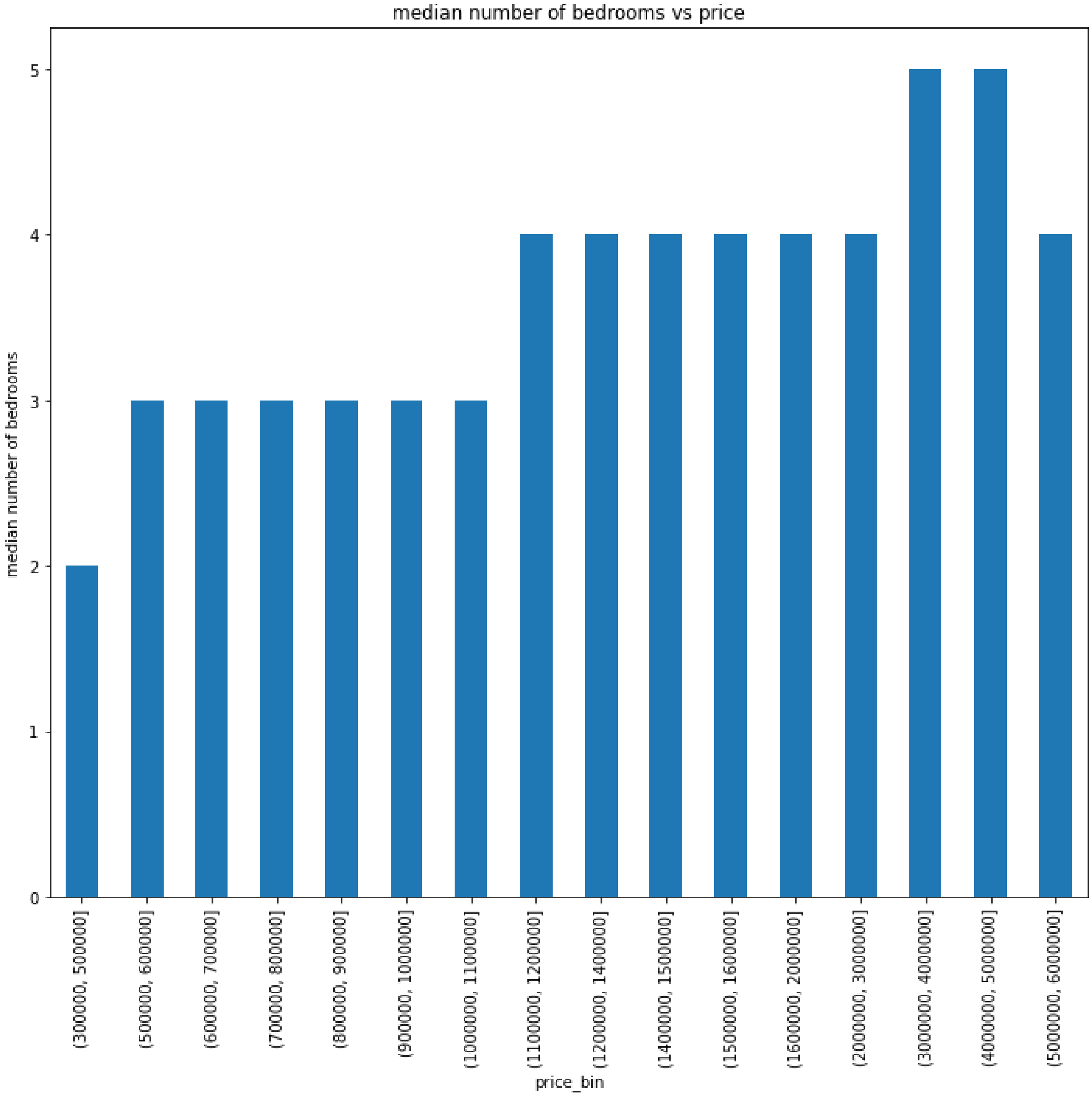


Fig 13a: Price_bin vs median bedrooms (single family)

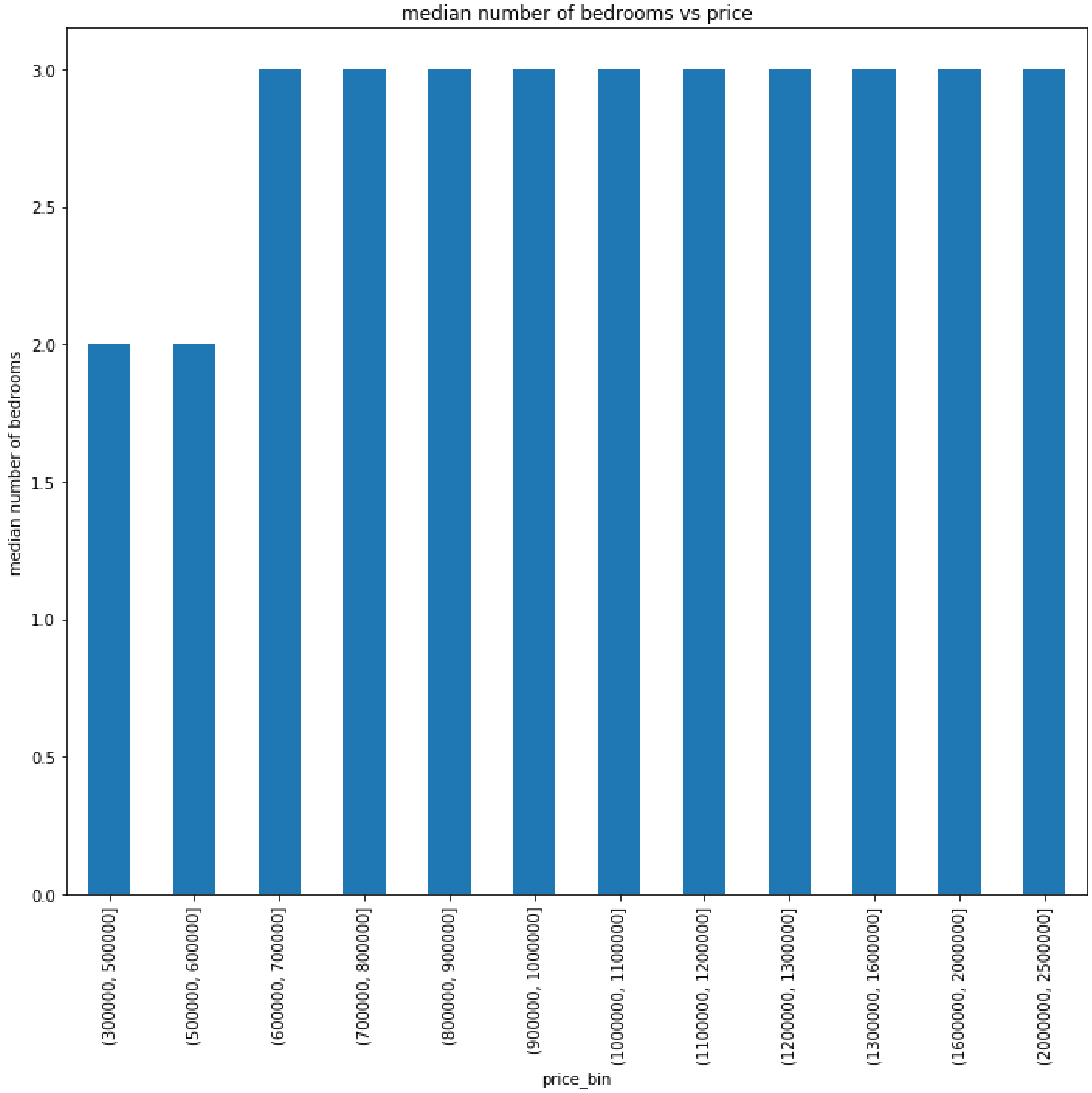


Fig 13b: Price_bin vs median bedrooms (townhomes)

How many number of bathrooms in various price bin?

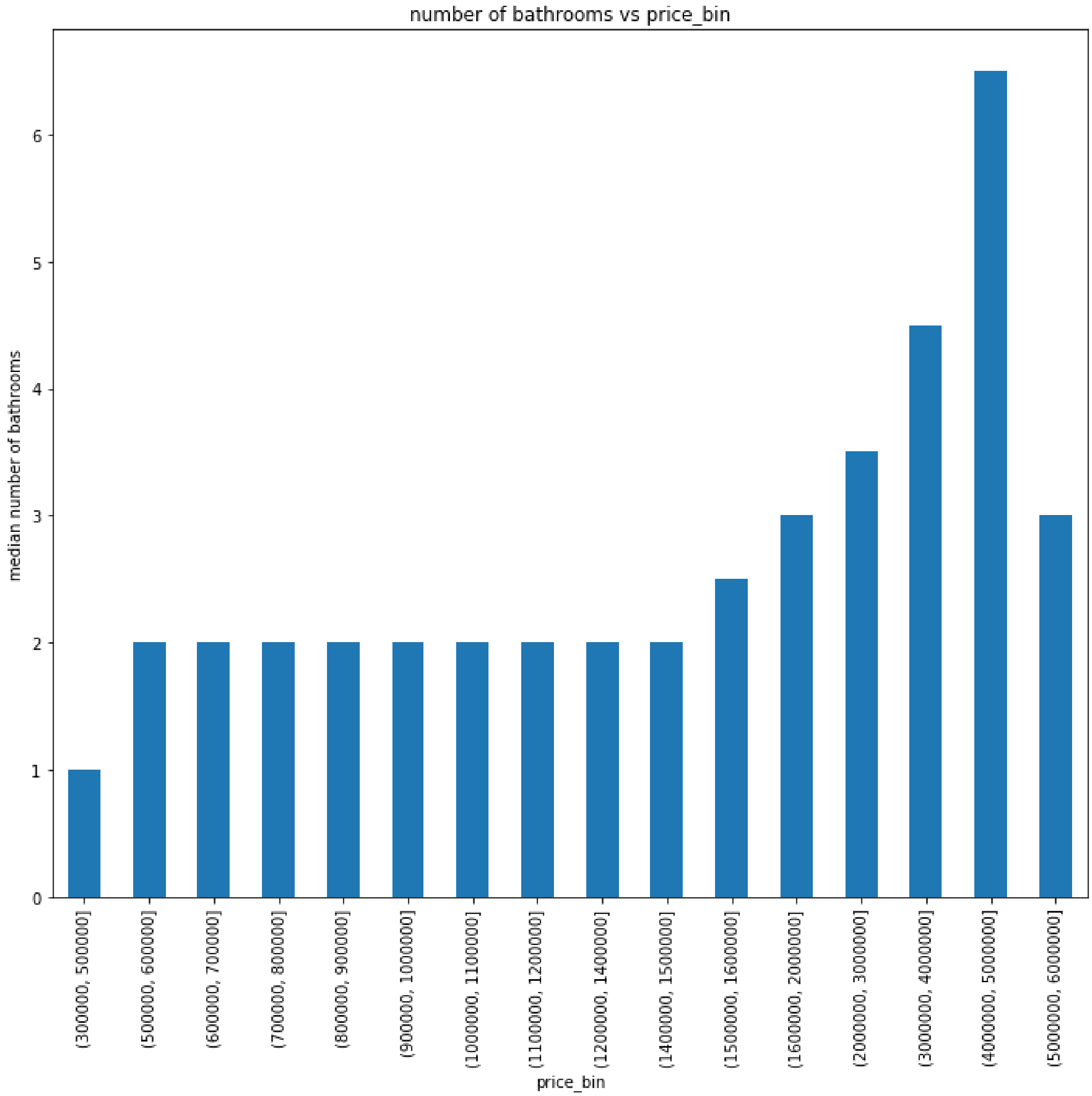


Fig 14a: Price_bin vs median bathrooms (single family)

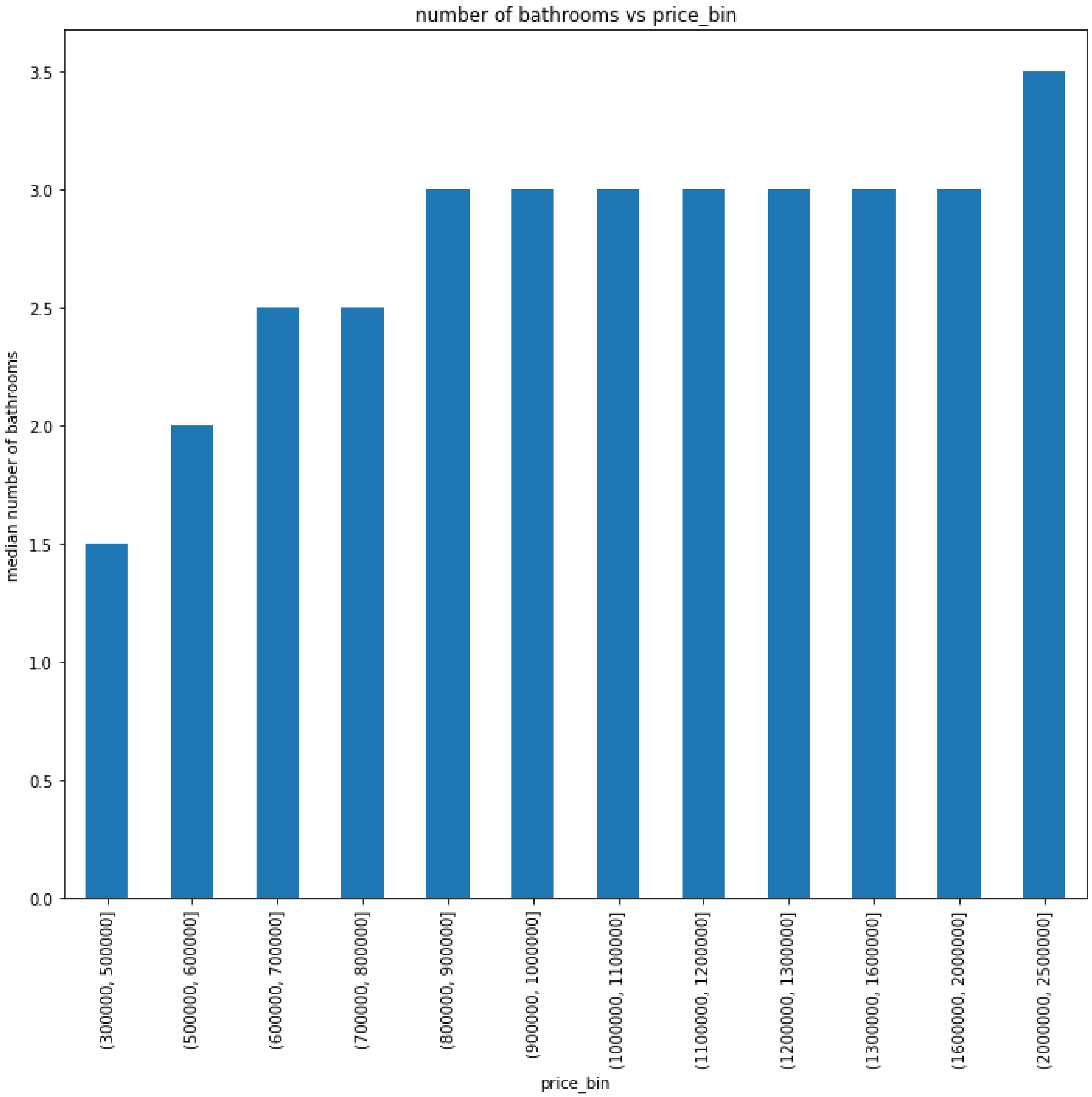


Fig 14b: Price_bin vs median bathrooms (townhomes)

Which median sqft is common within various price bin?

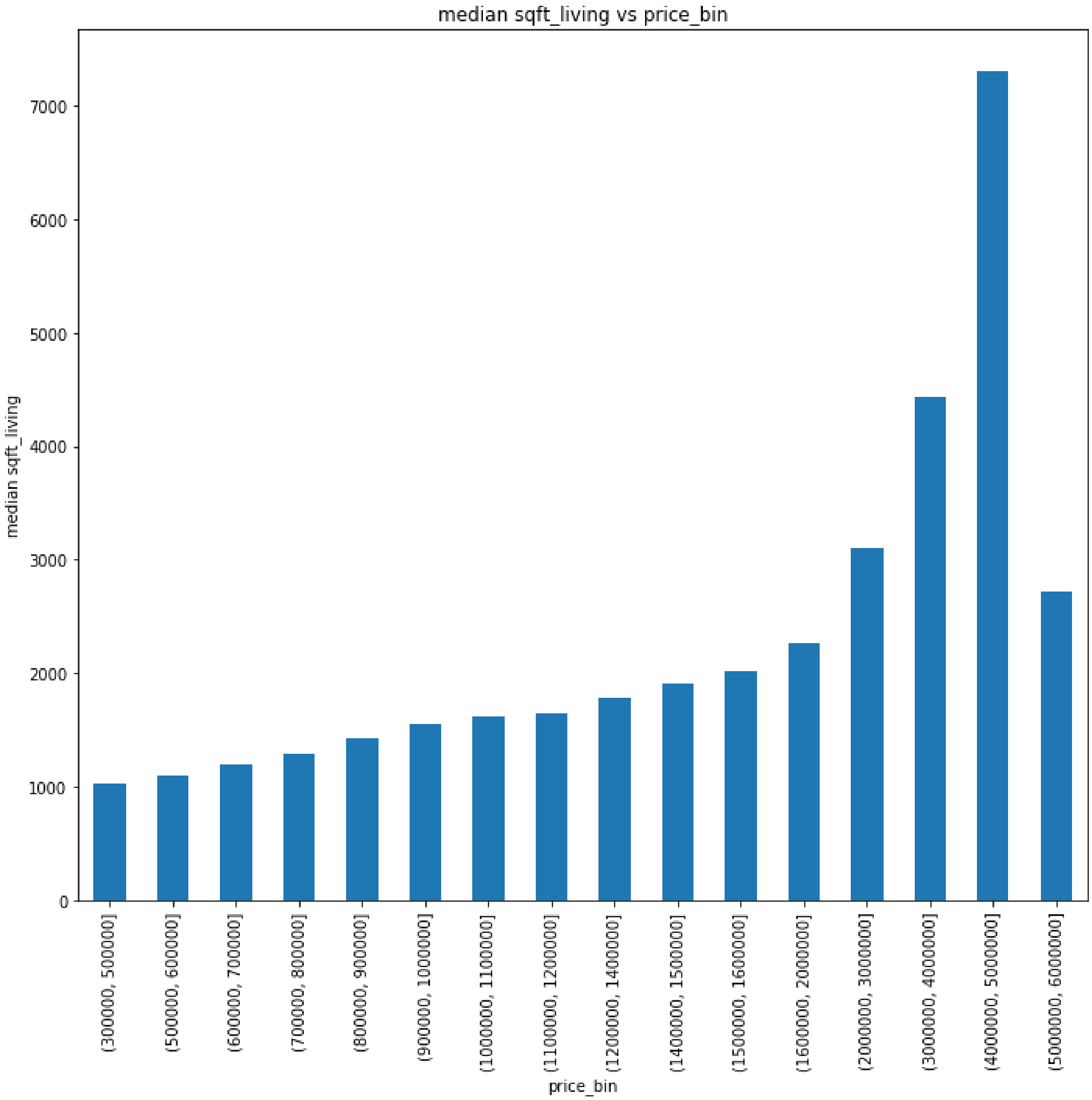


Fig 15a: Price_bin vs median sqft (single family)

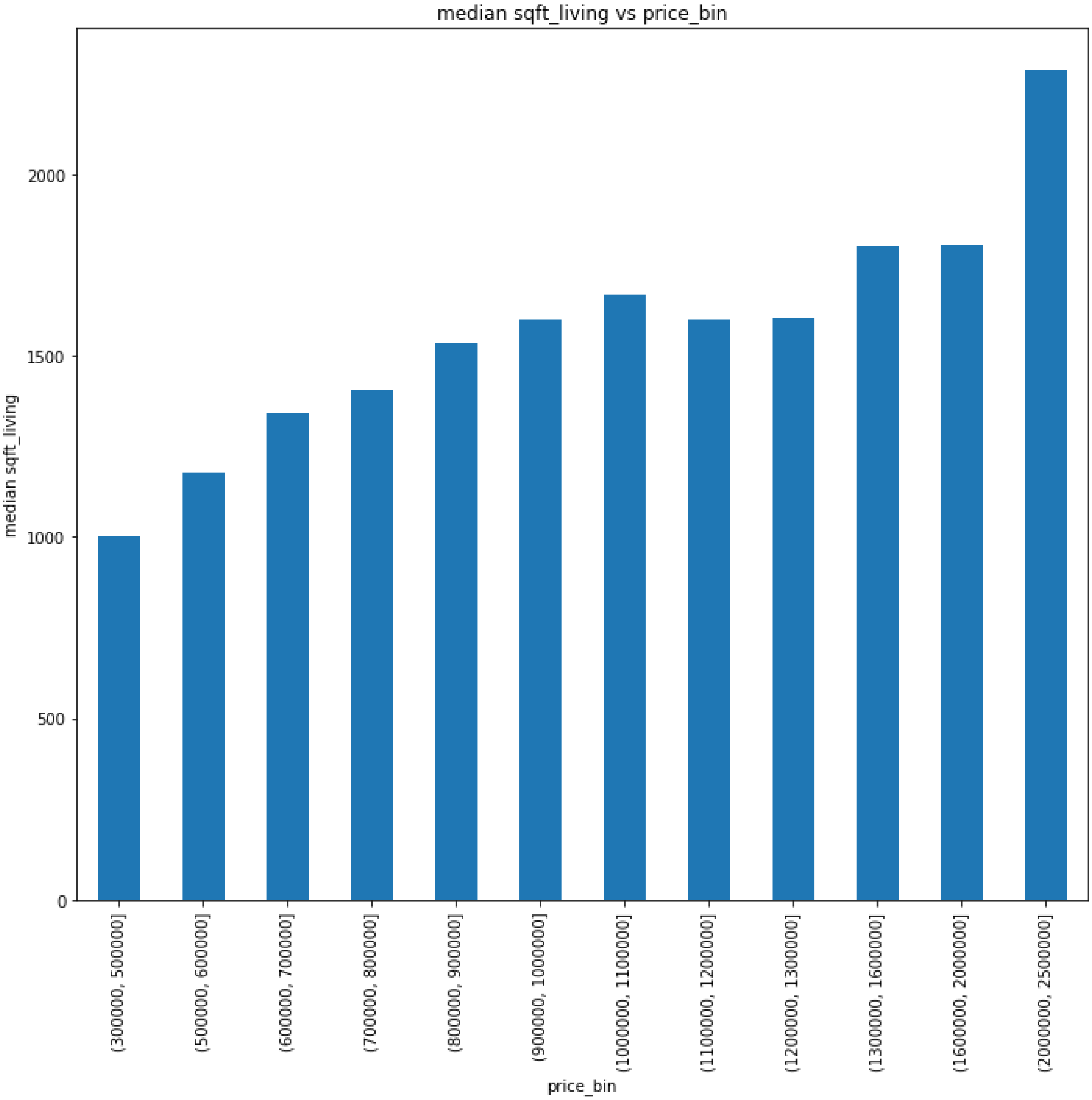


Fig 15b: Price_bin vs median sqft (townhomes)

Which price range homes are in liquefaction zone?

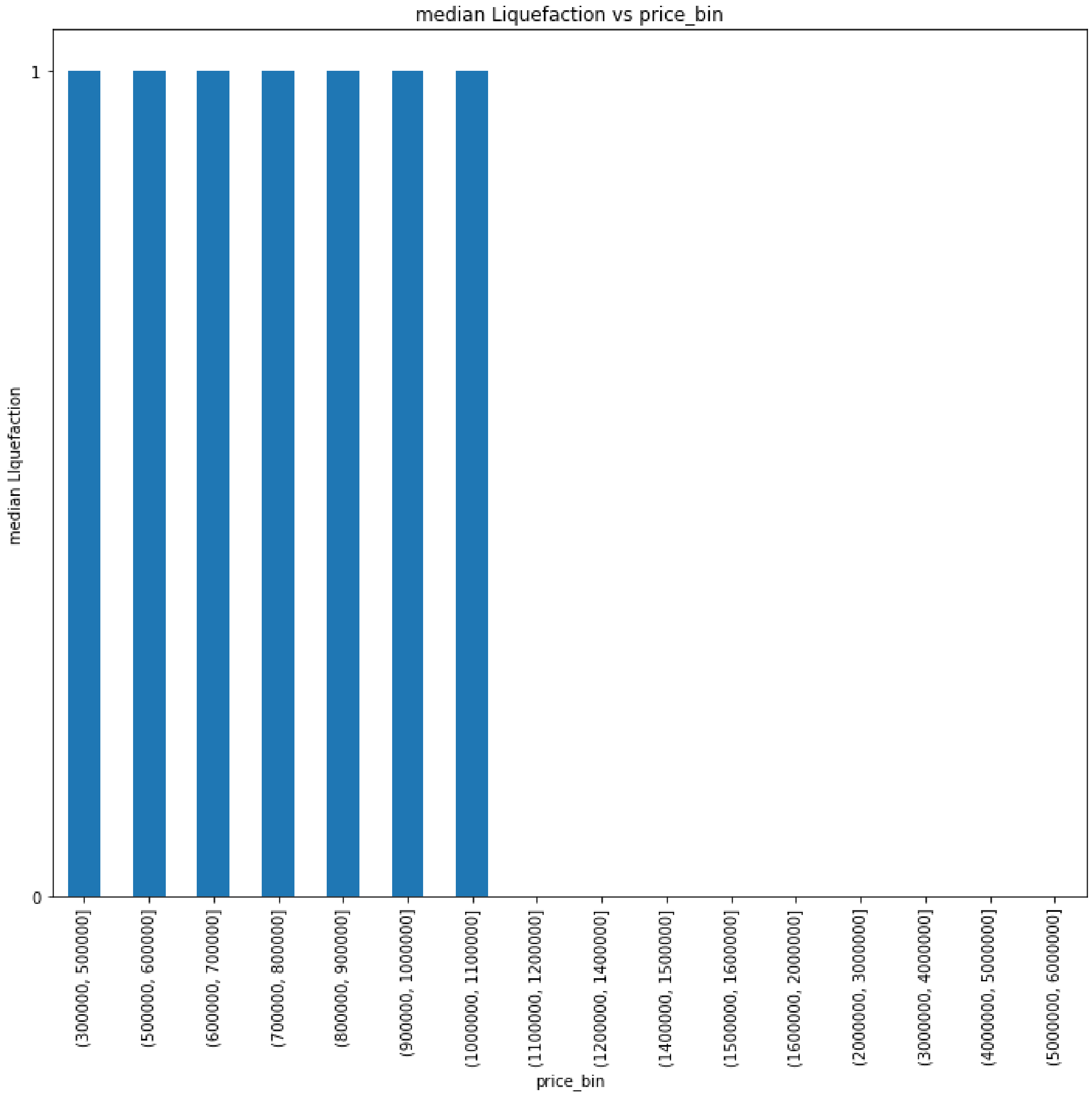


Fig 16a: Price_bin vs liquefaction (single family)

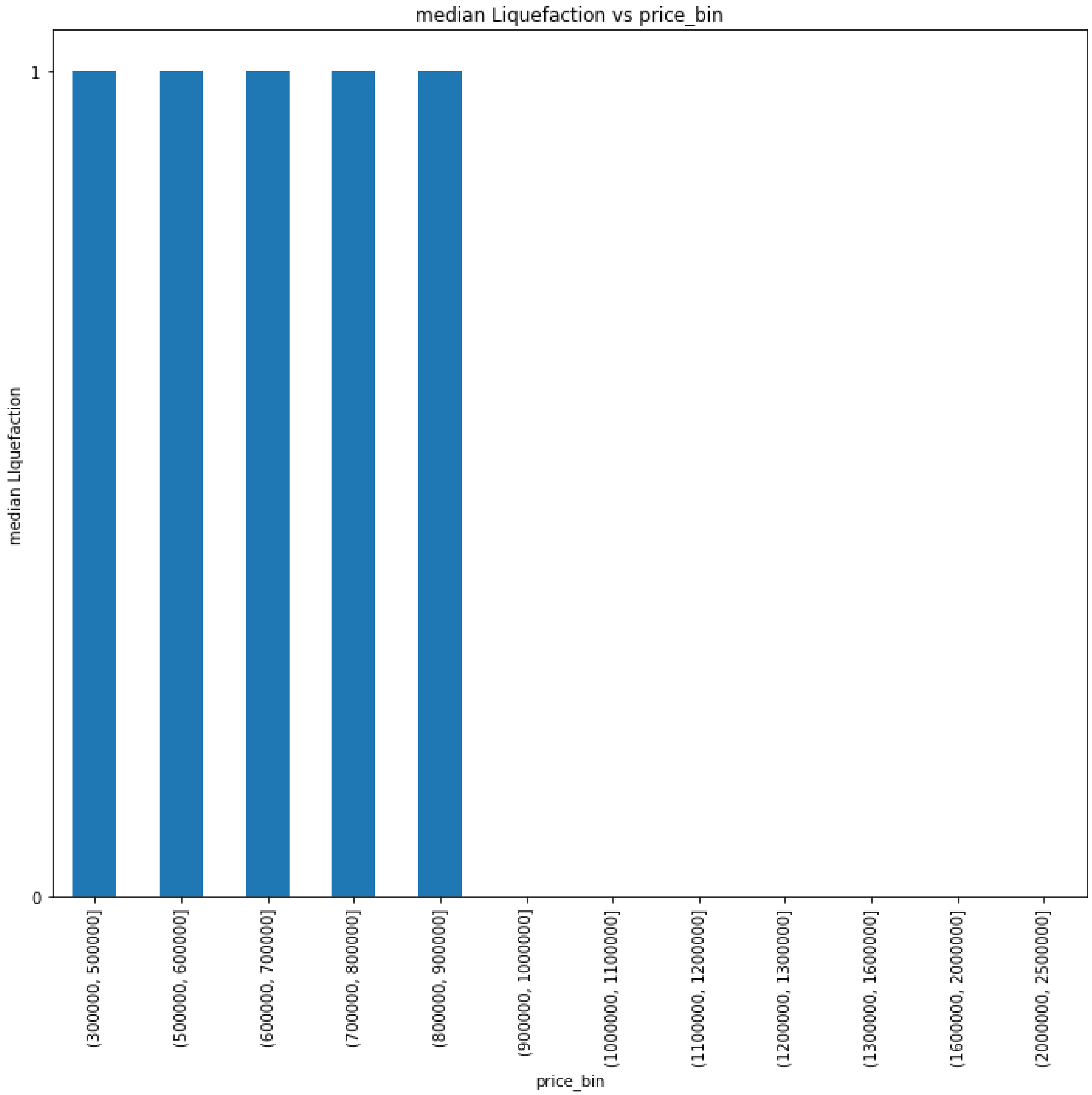


Fig 16b: Price_bin vs liquefaction (townhomes)

Which price range homes are in landslide zone?

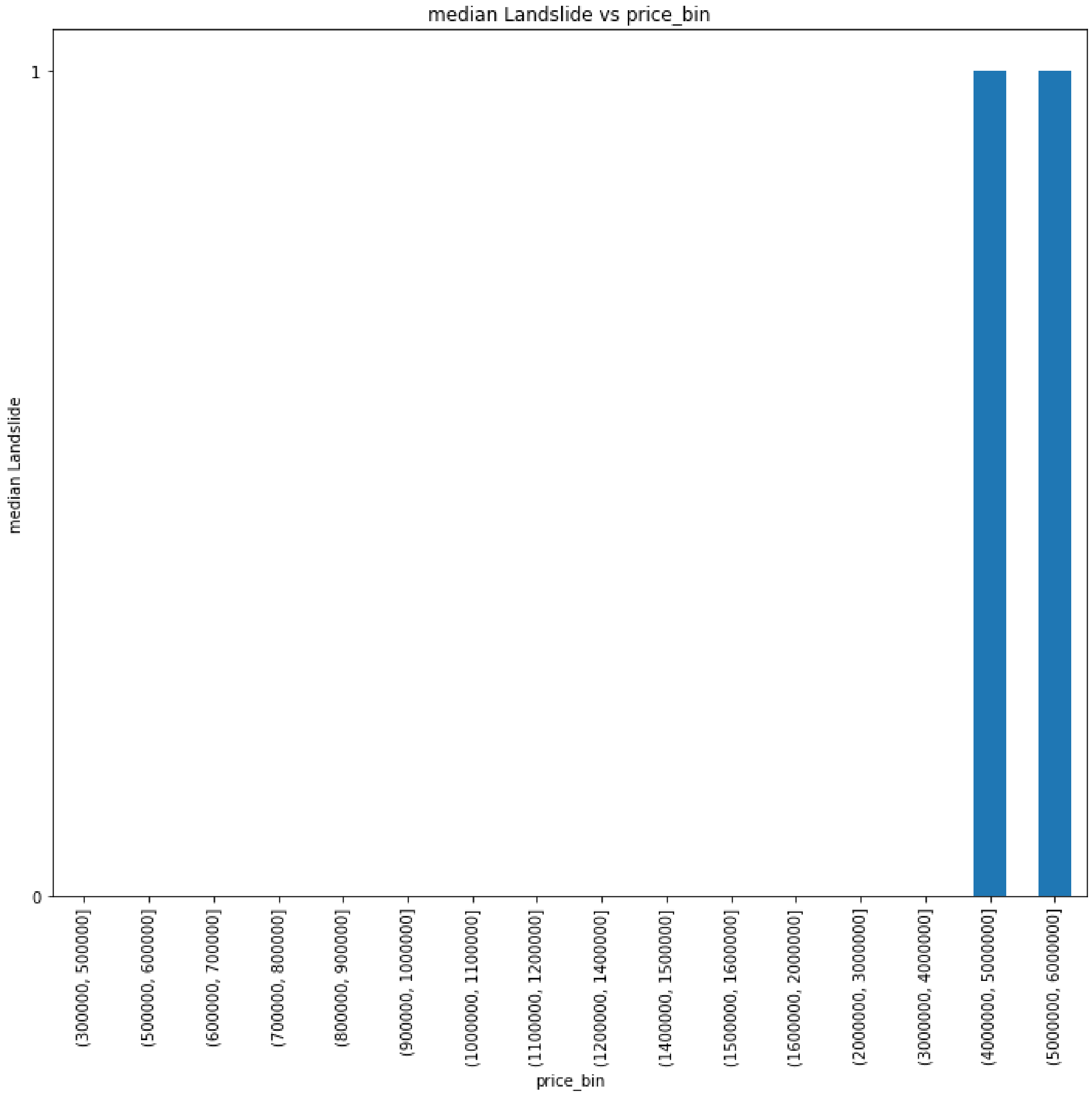


Fig 17: Price_bin vs landslide (single family)

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

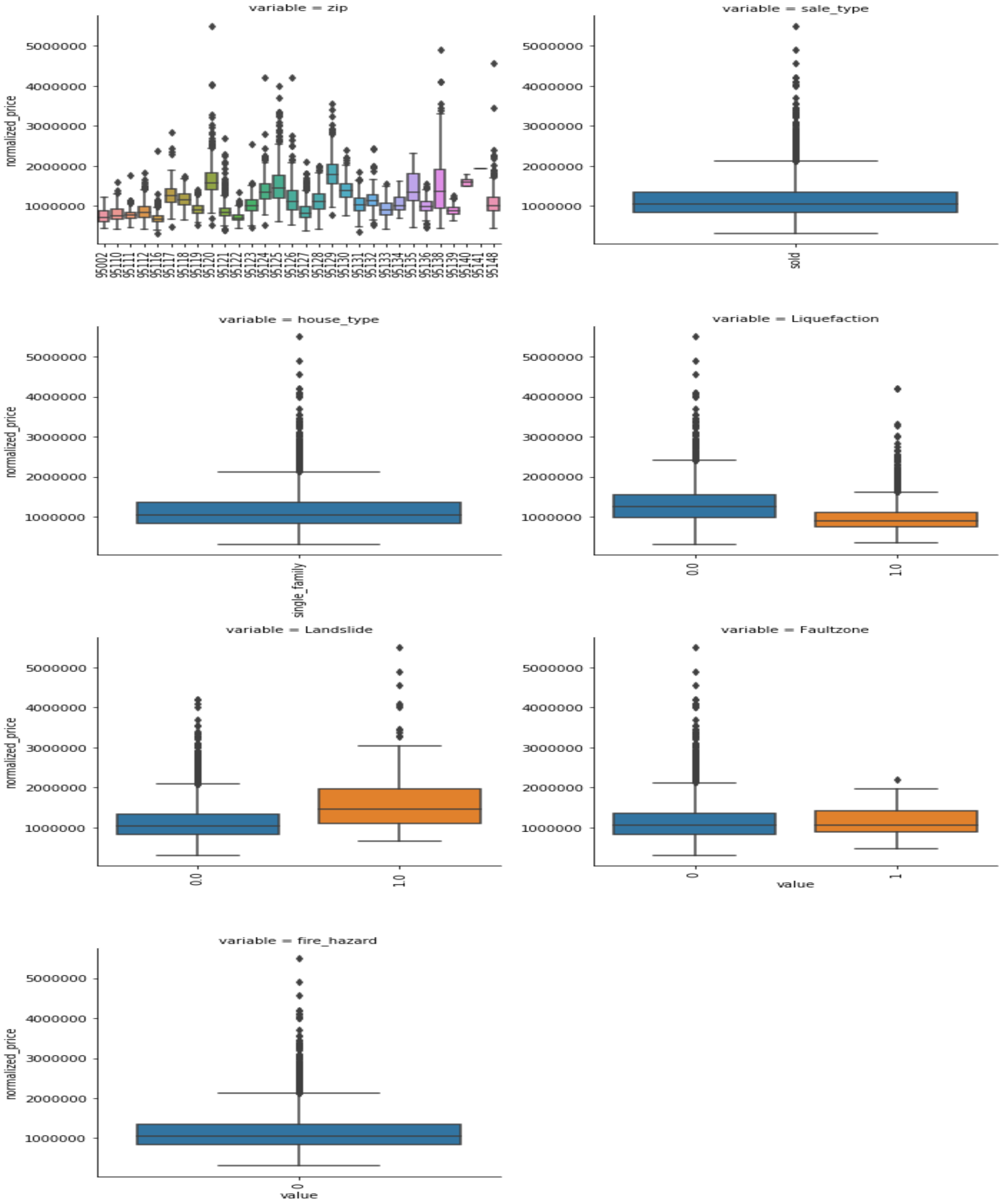


Fig 18a: Price distribution box plot (single family)

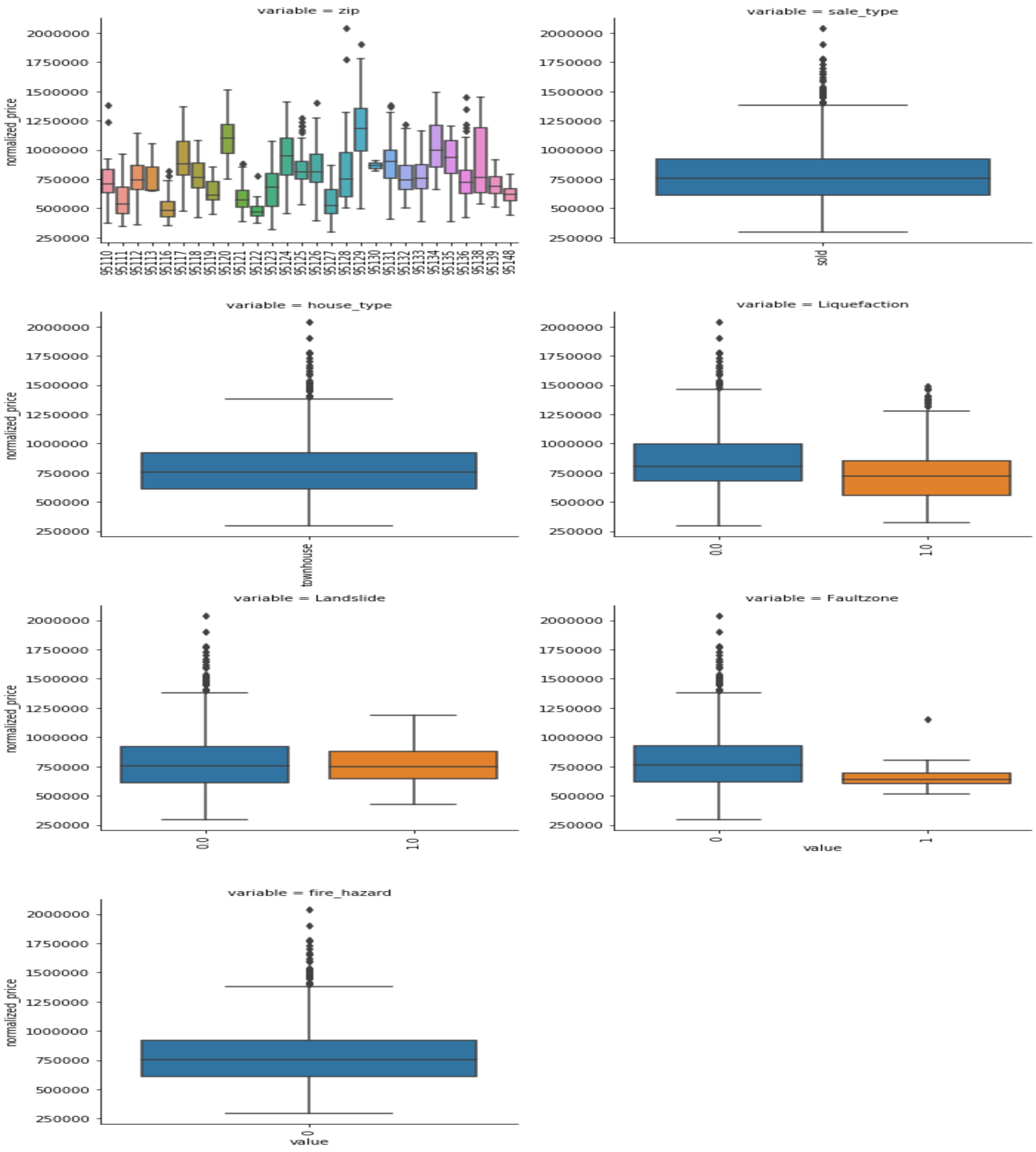


Fig 18b: Price distribution box plot (townhomes)

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?

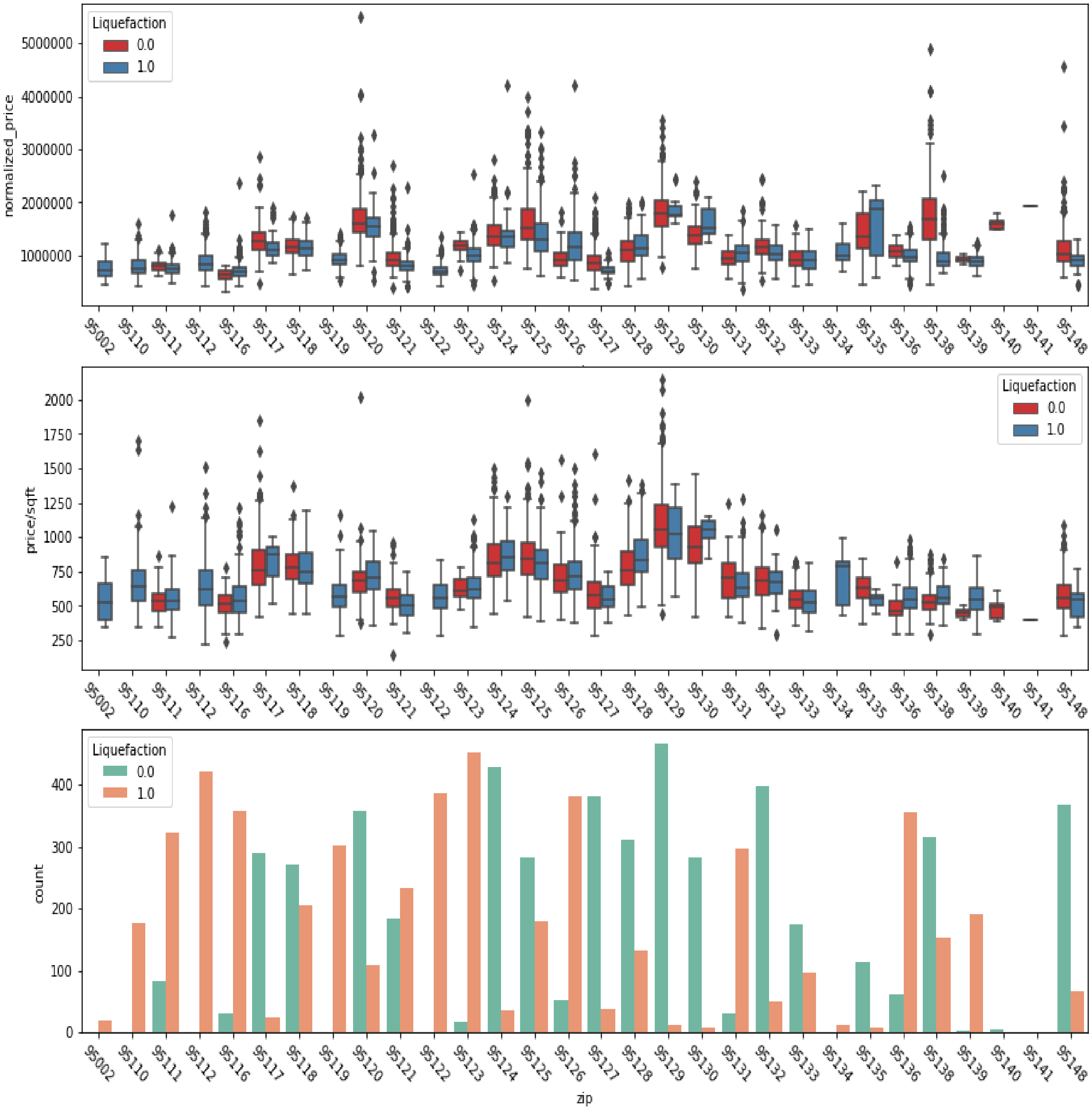


Fig 19a: Price distribution box plot (single family)

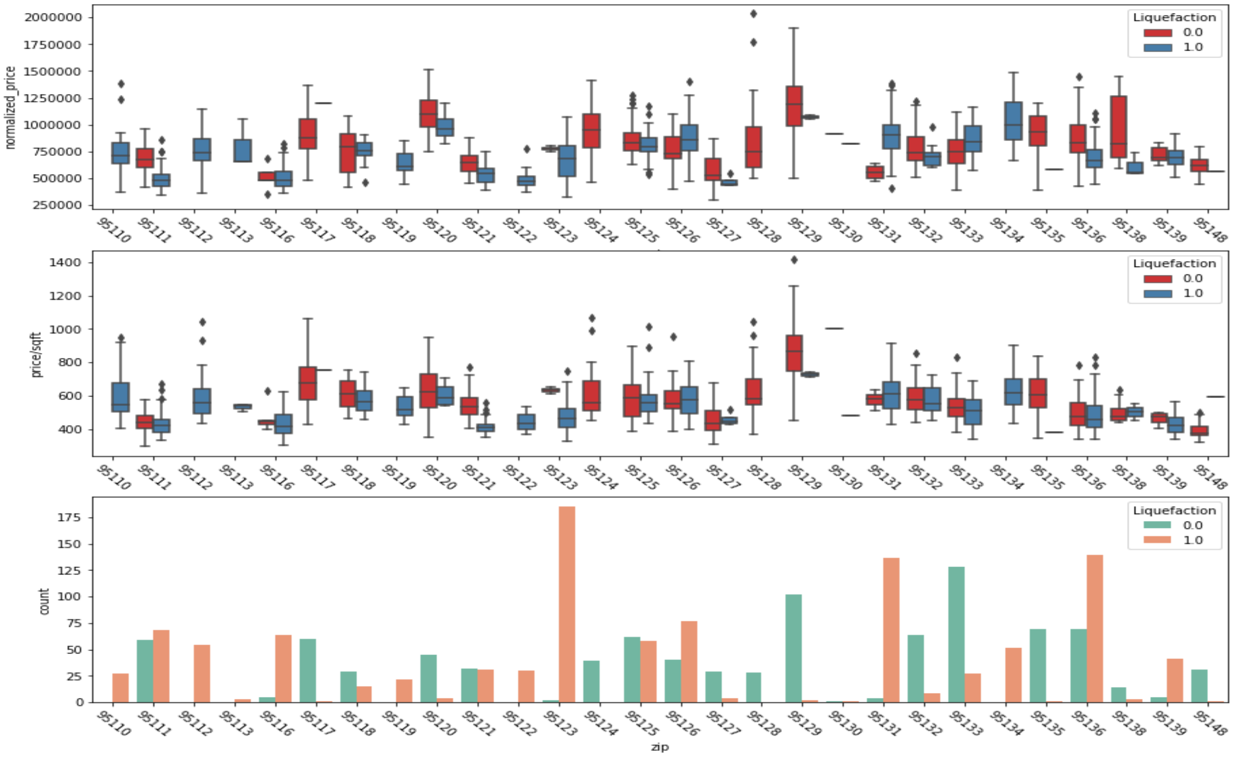


Fig 19b: Price distribution box plot (townhomes)

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?

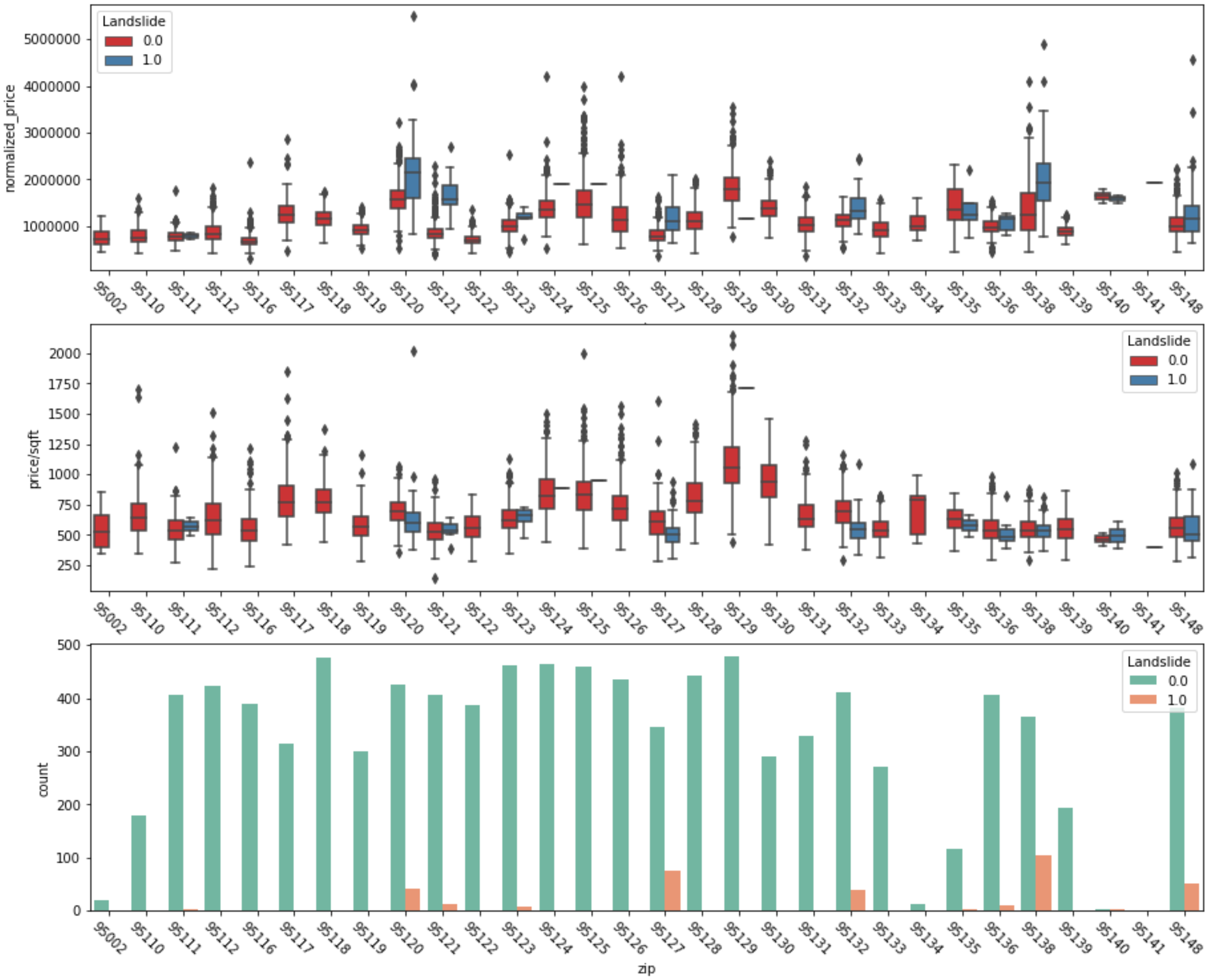


Fig 20a: Price distribution box plot (single family)

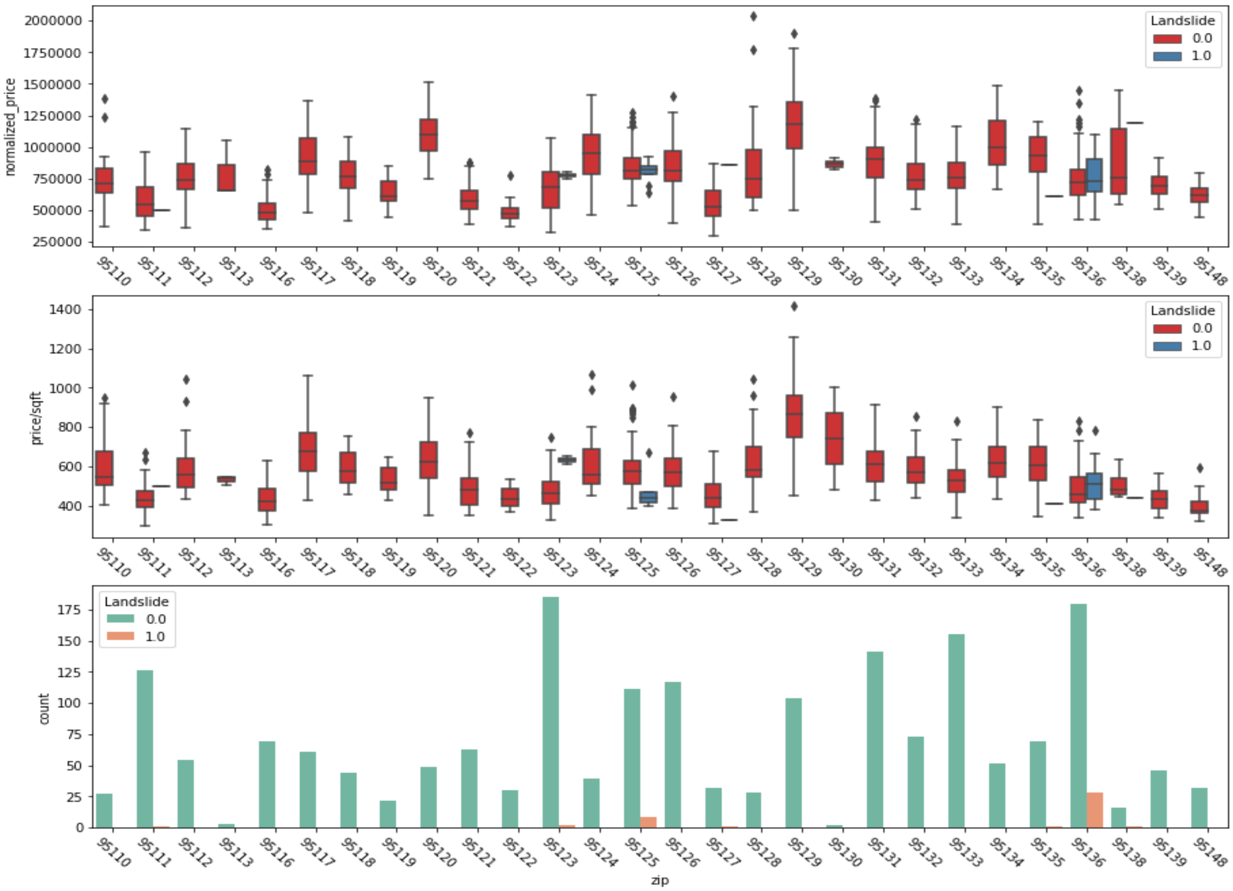


Fig 20b: Price distribution box plot (townhomes)

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?

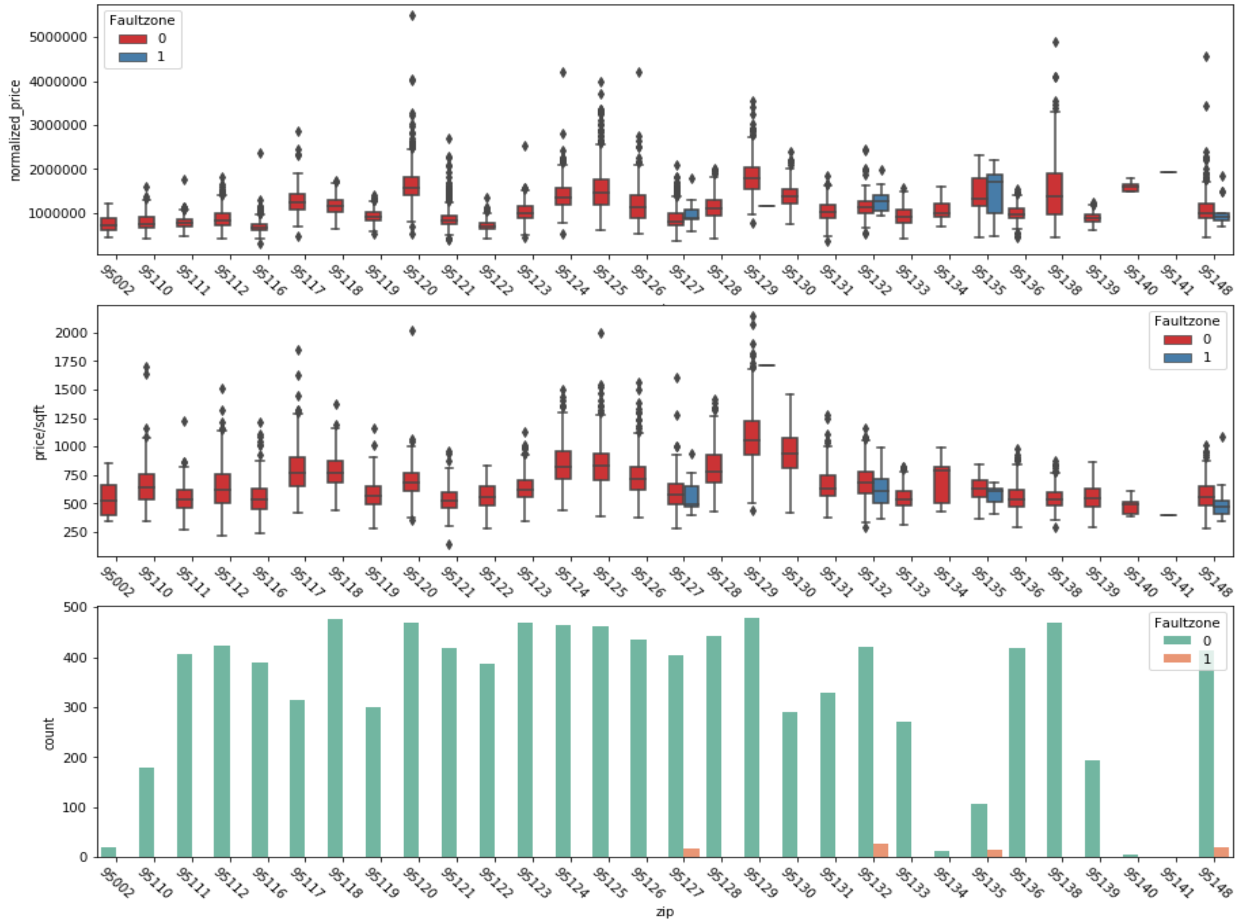


Fig 21a: Price distribution box plot (single family)

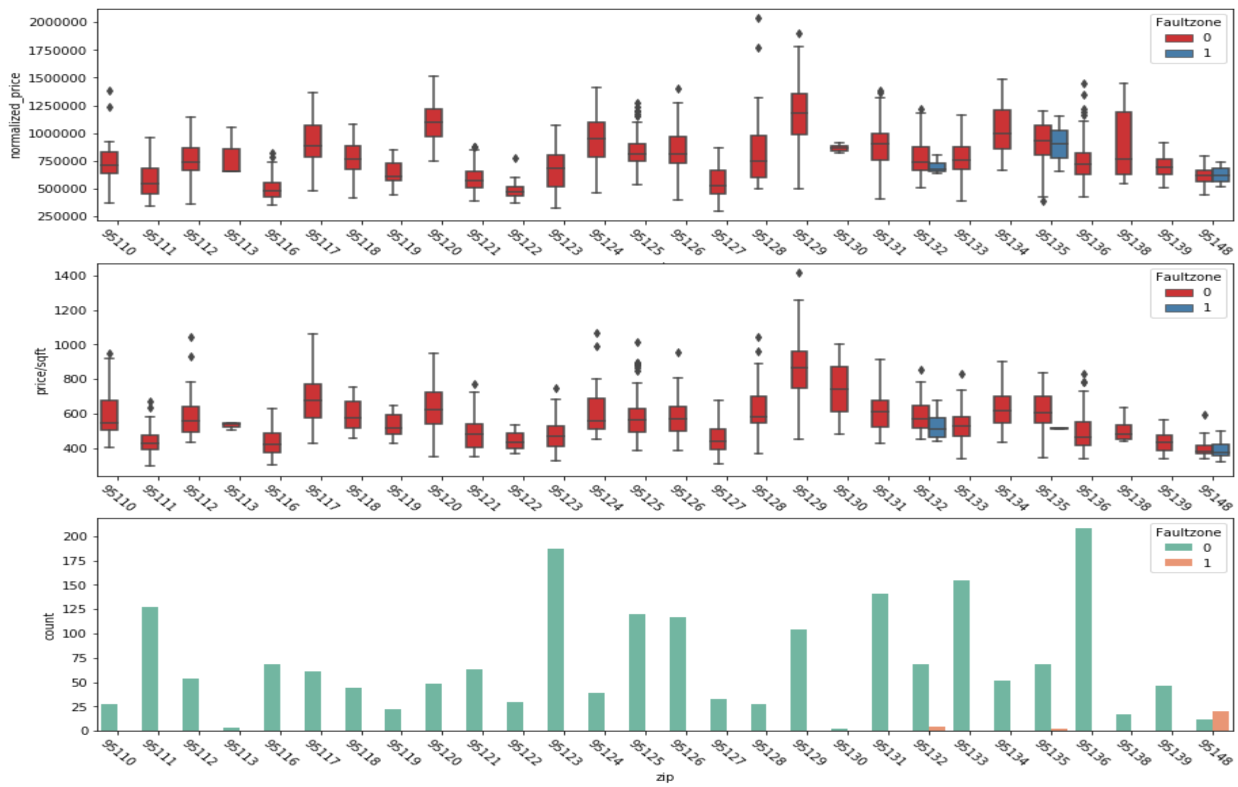


Fig 21b: Price distribution box plot (townhomes)

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?

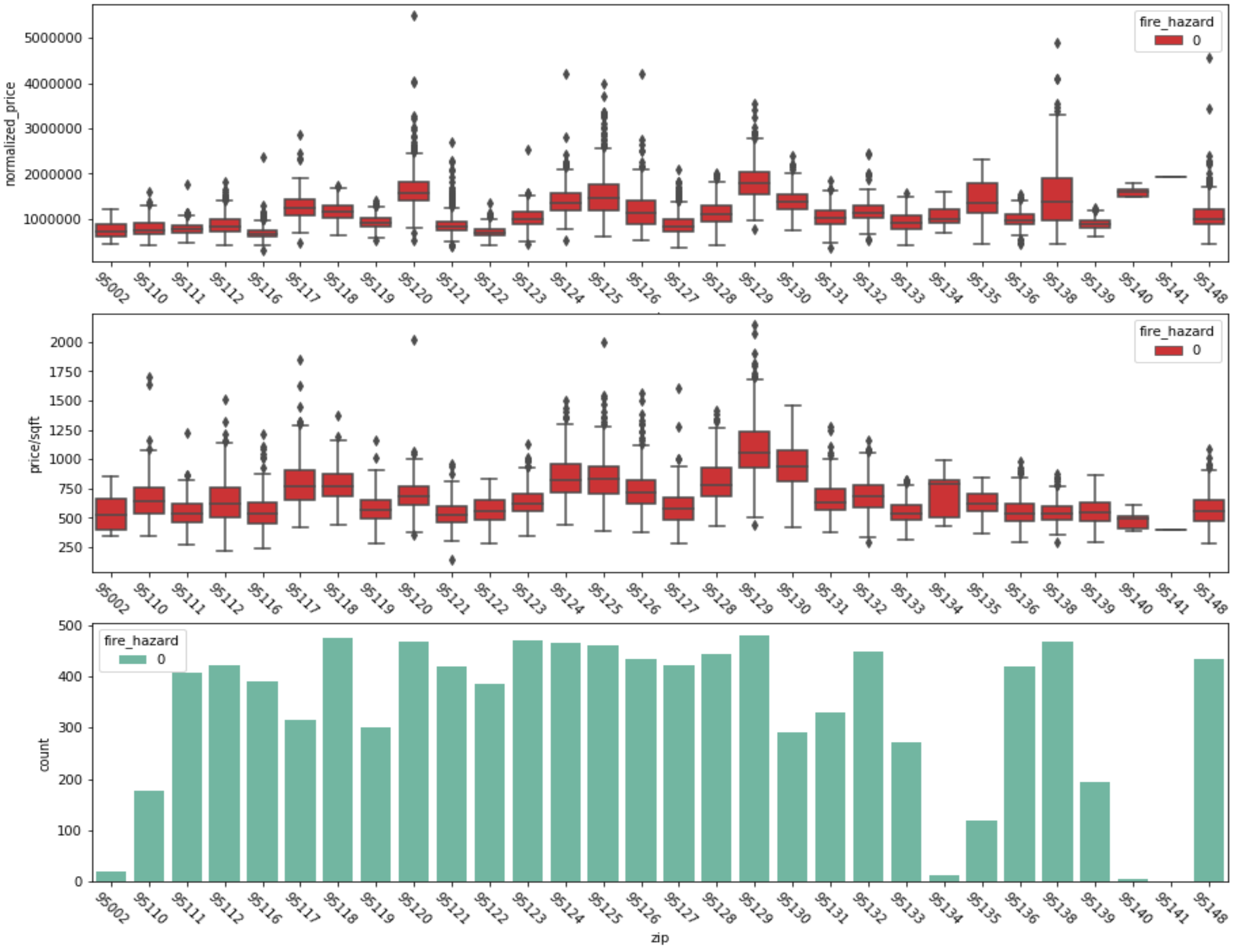


Fig 22a: Price distribution box plot (single family)

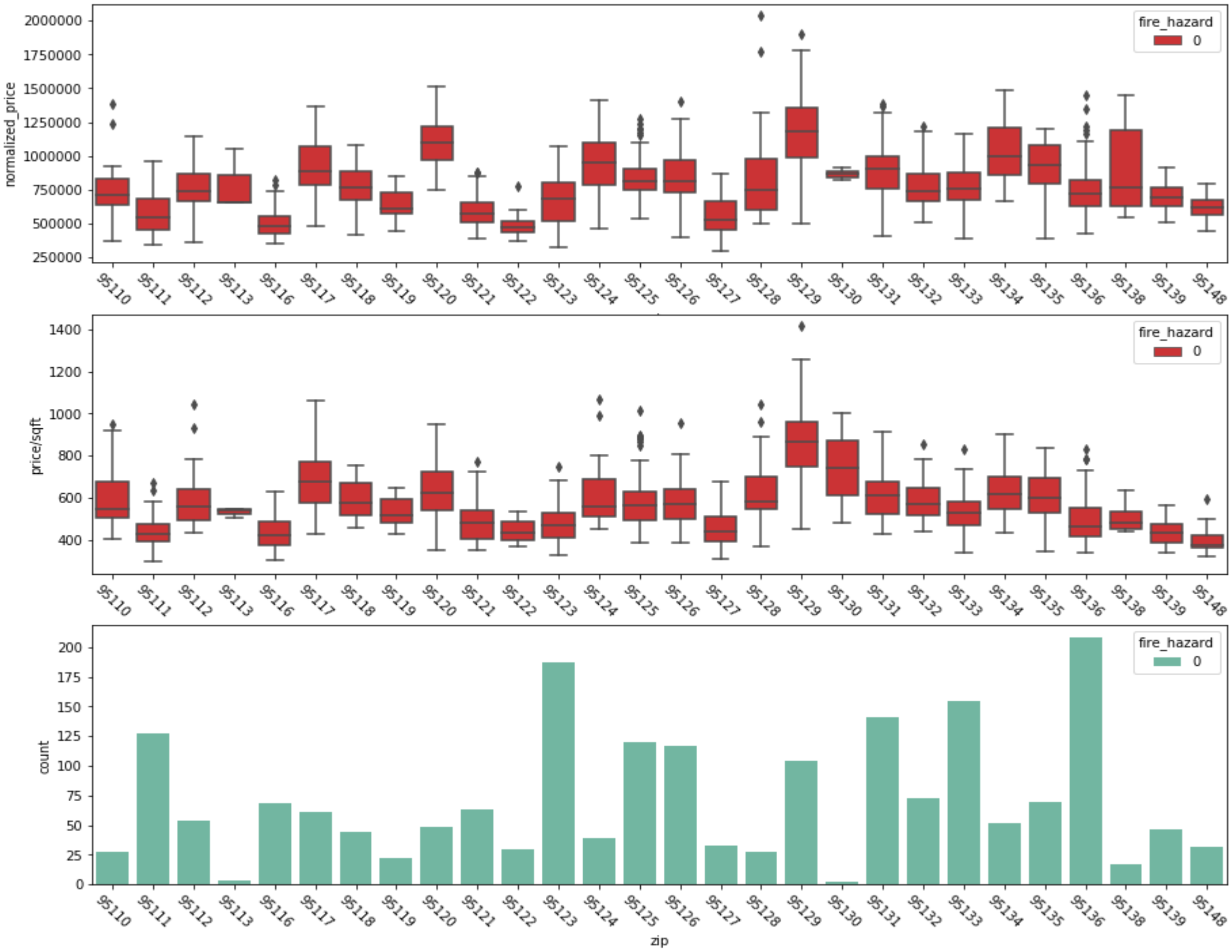


Fig 22b: Price distribution box plot (townhomes)

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?

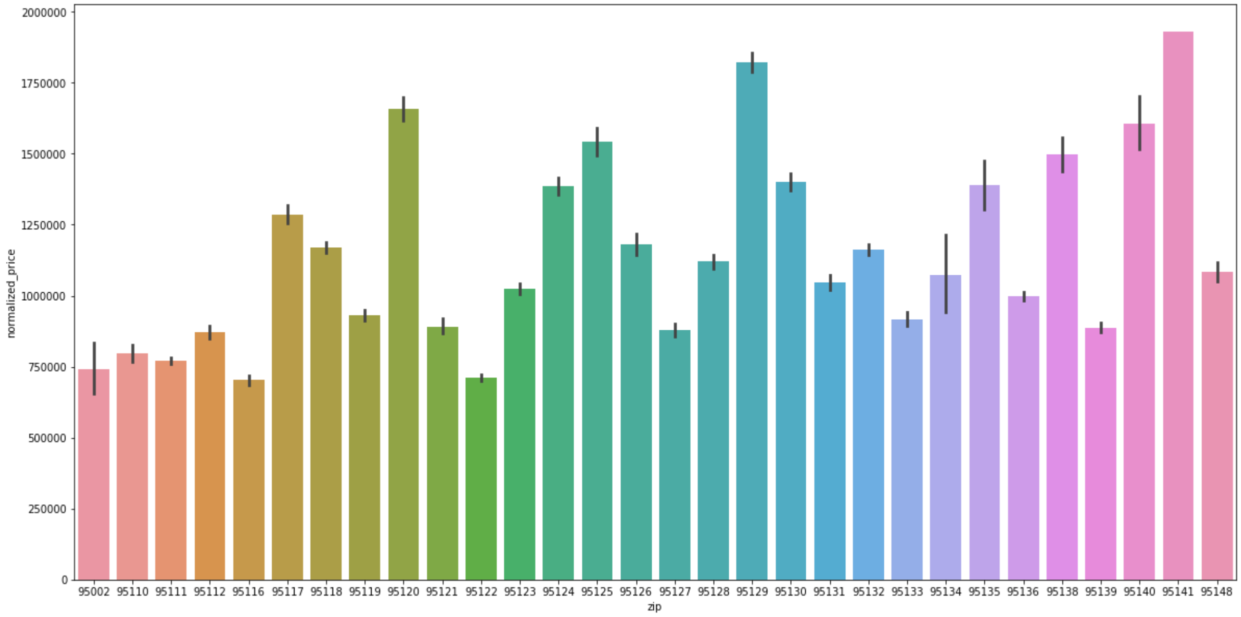


Fig 23a: Price distribution bar plot (single family)

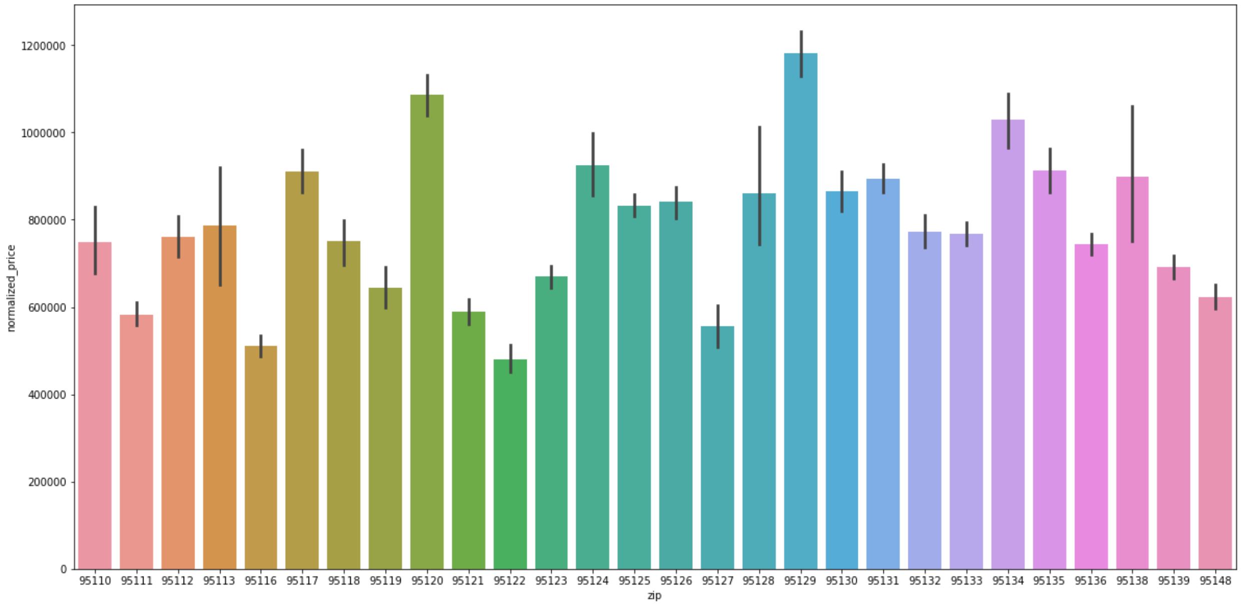


Fig 23a: Price distribution bar plot (townhomes)

What is the distribution of price per year built?

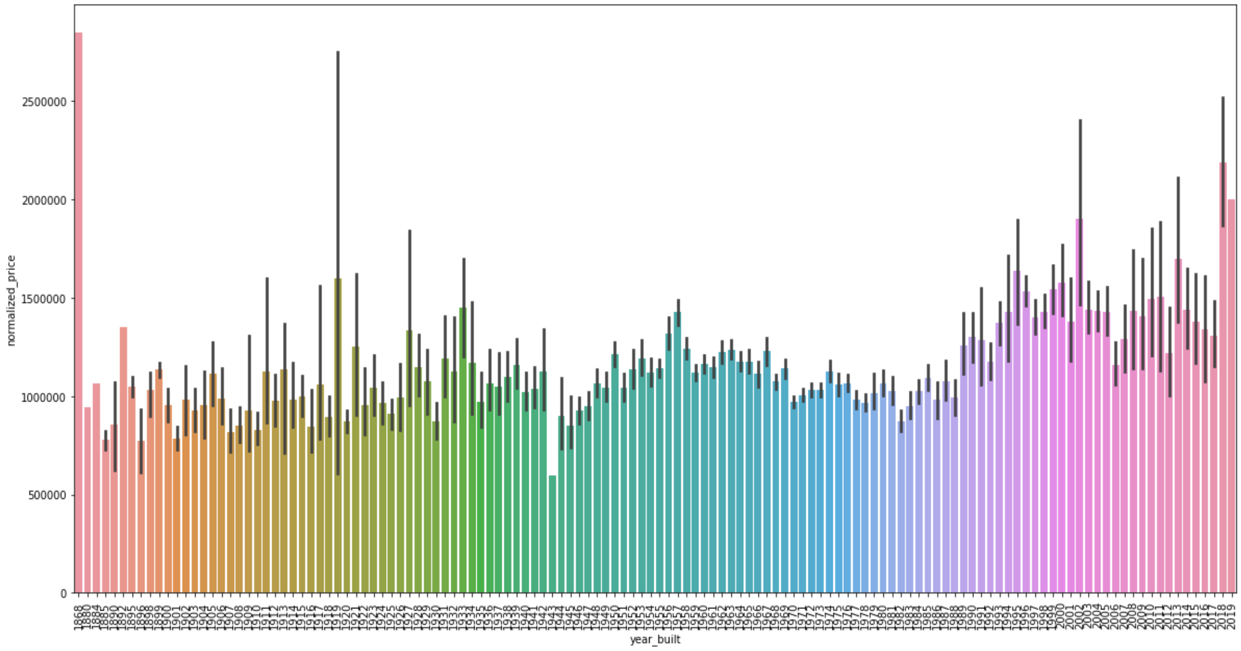


Fig 24a: Price distribution bar plot (single family)

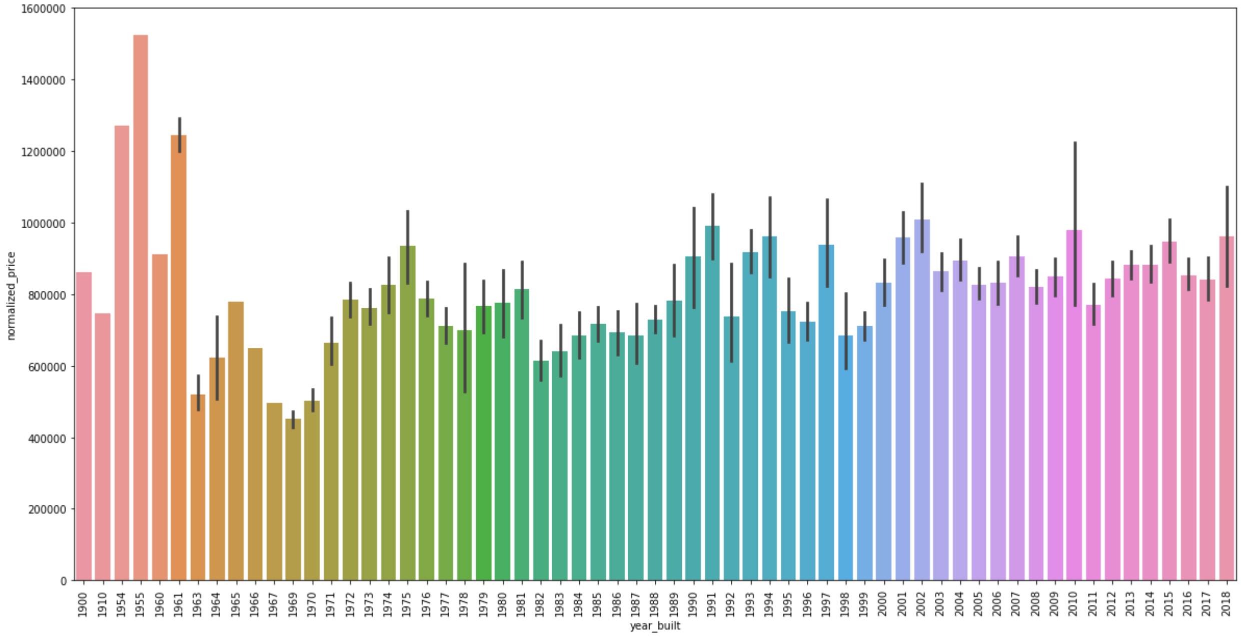
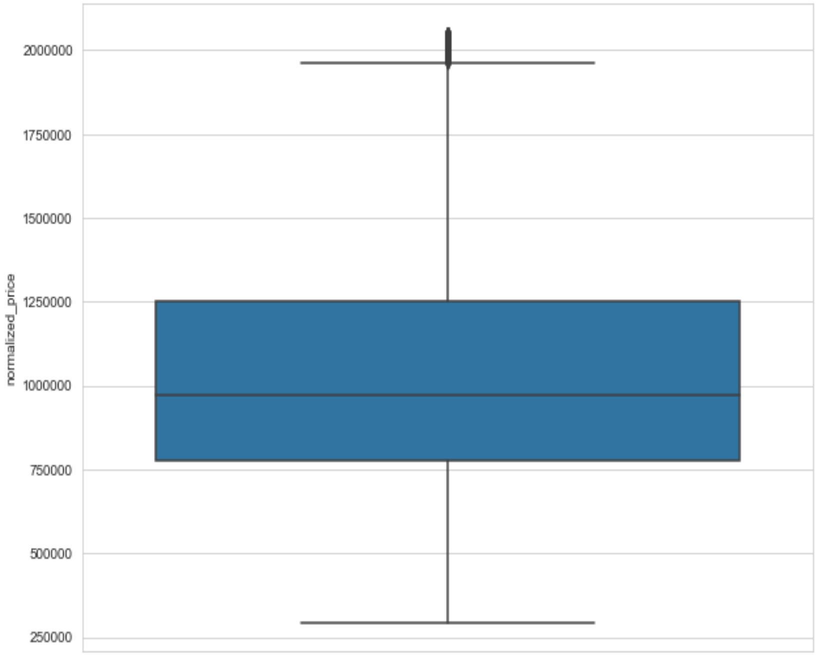
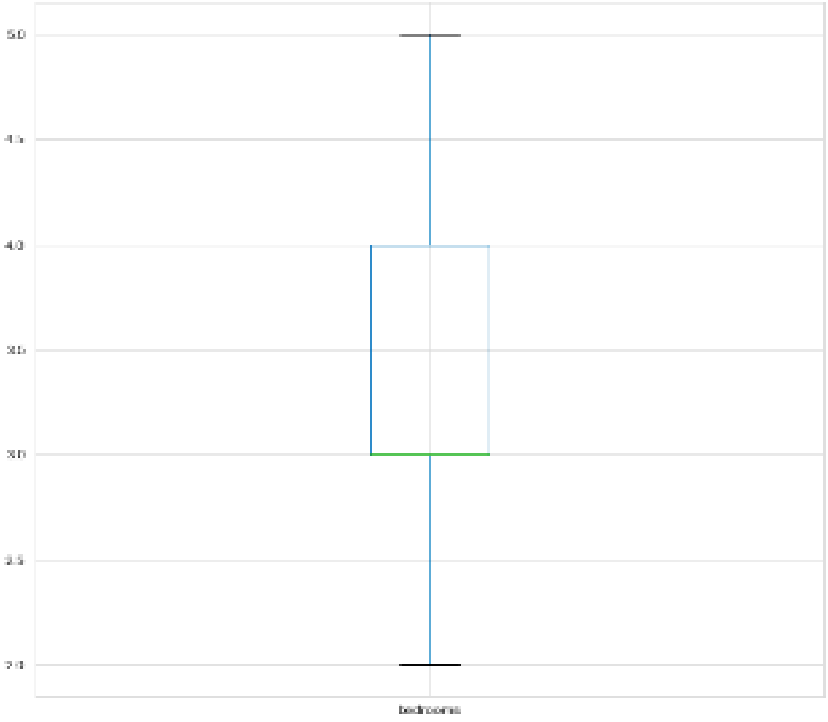


Fig 24b: Price distribution bar plot (townhomes)

Bar plot distribution for all features





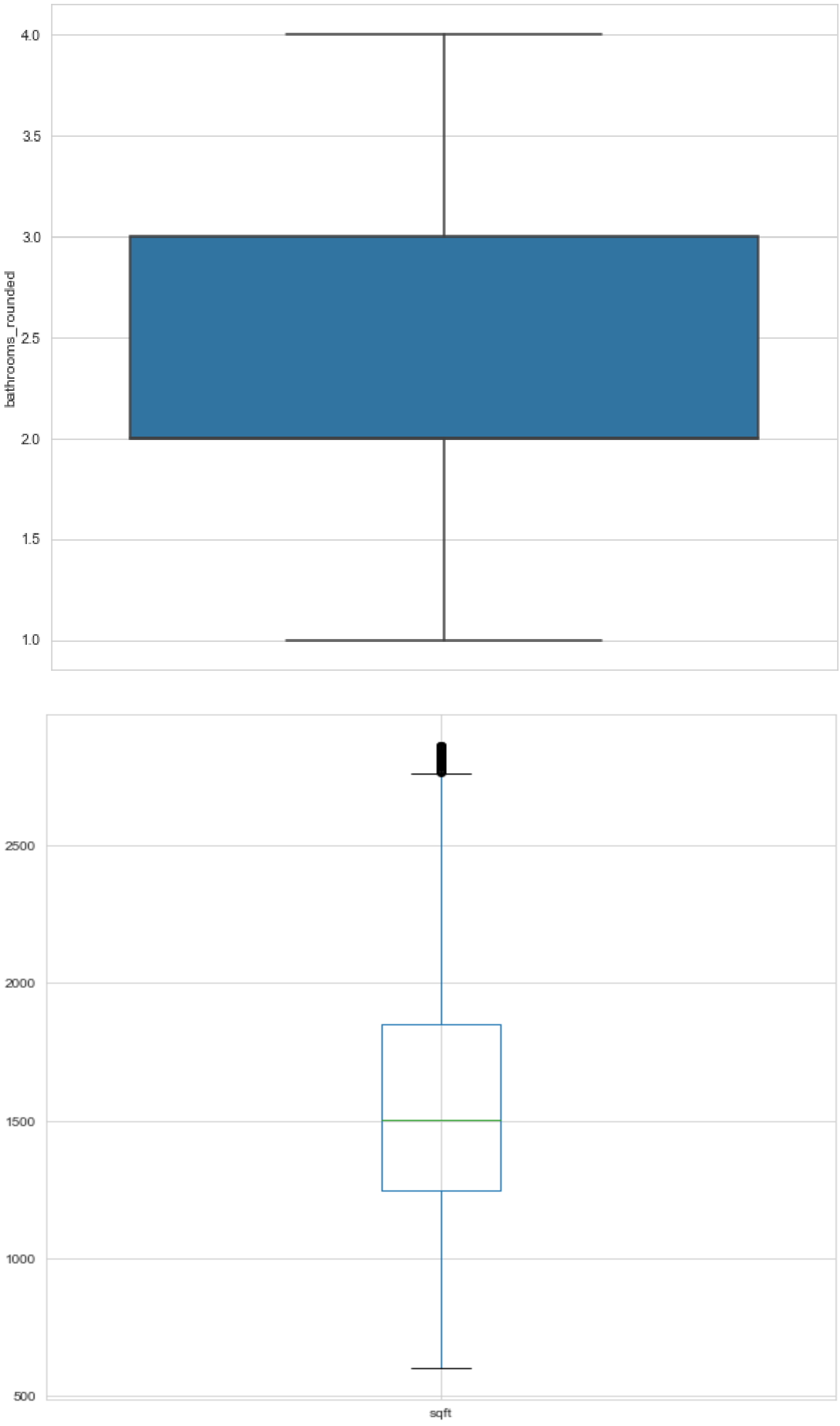


Fig 25: Bar plot distribution

Price distribution in geospatial frame

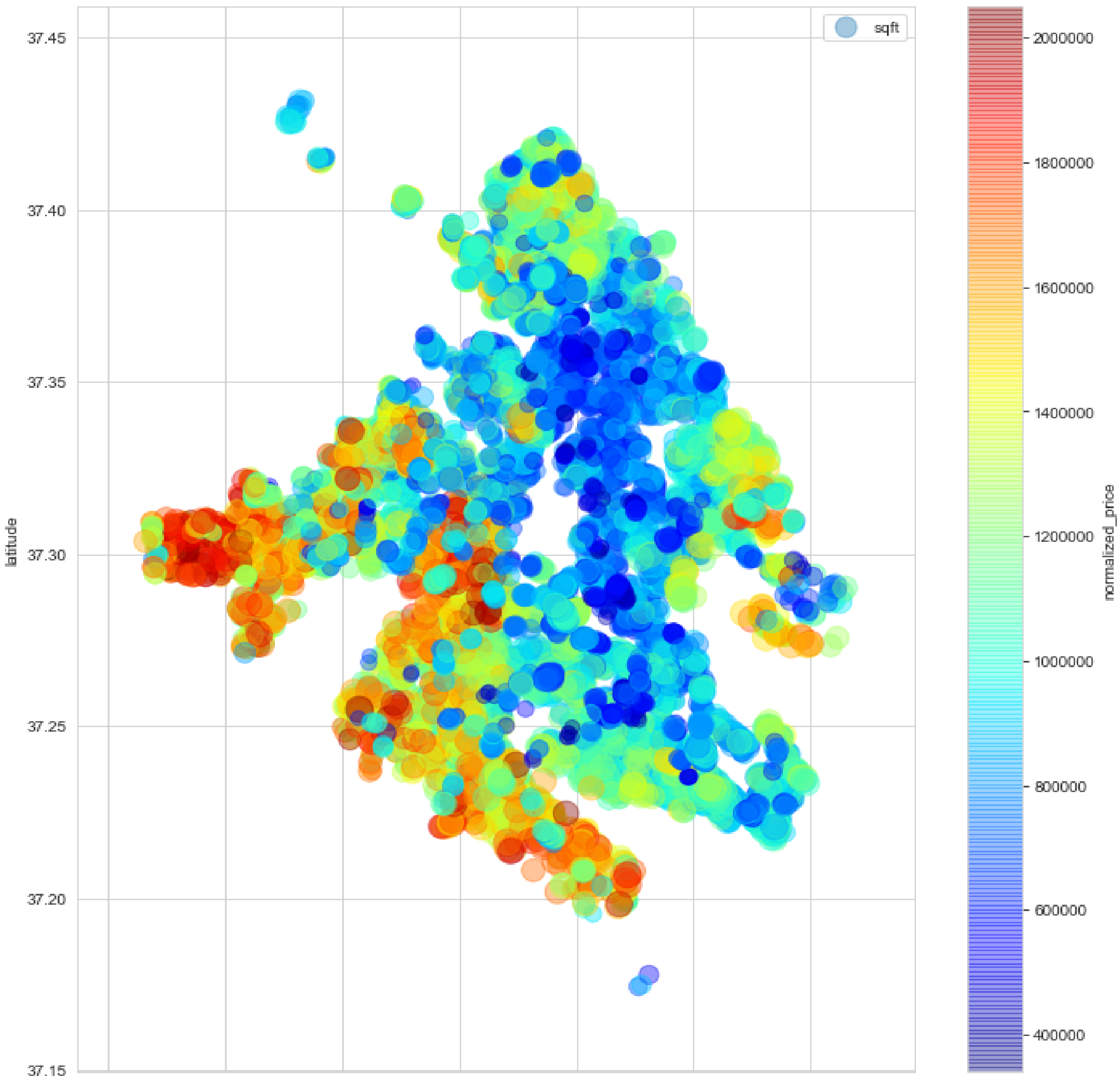
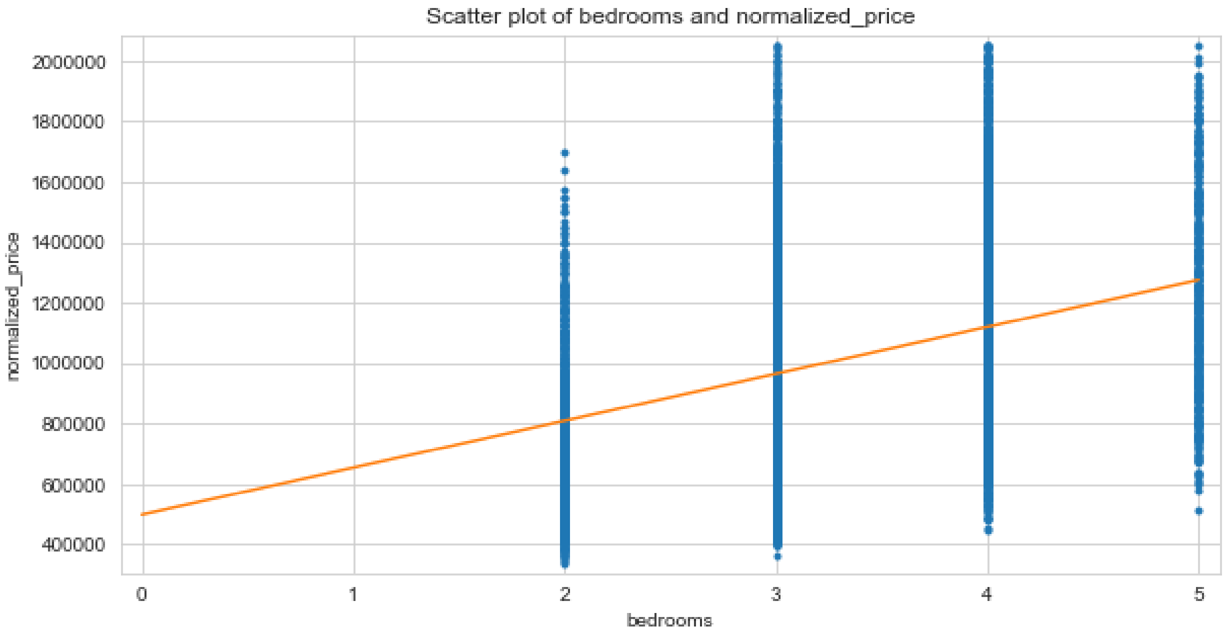


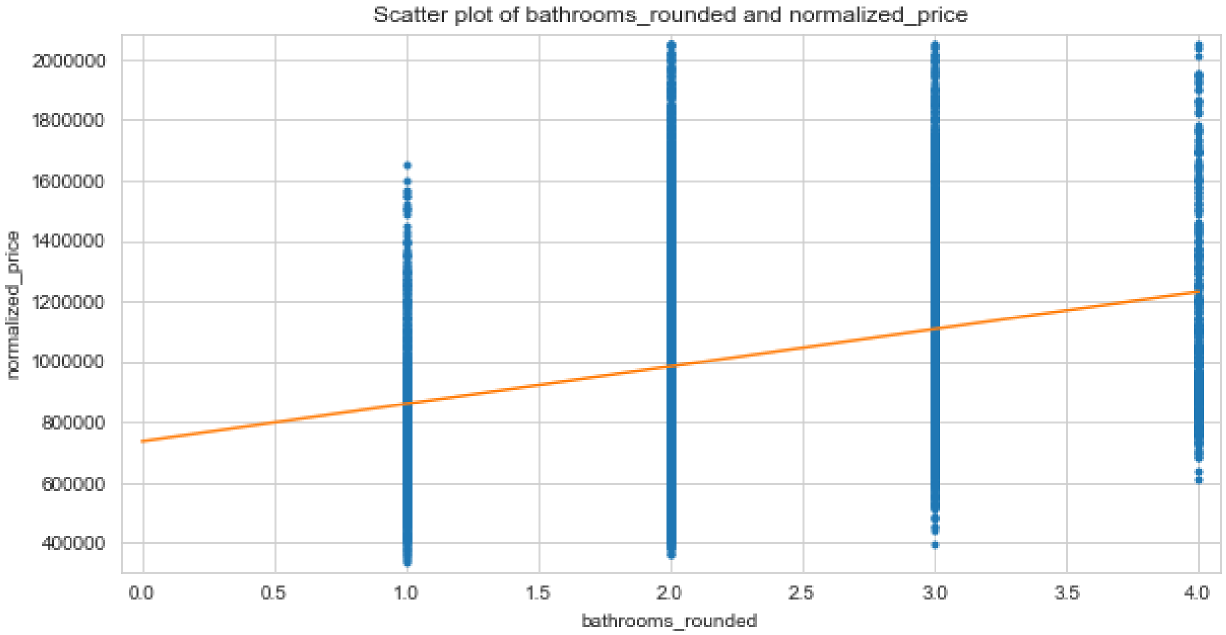
Fig 26: price distribution in geospatial frame

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.

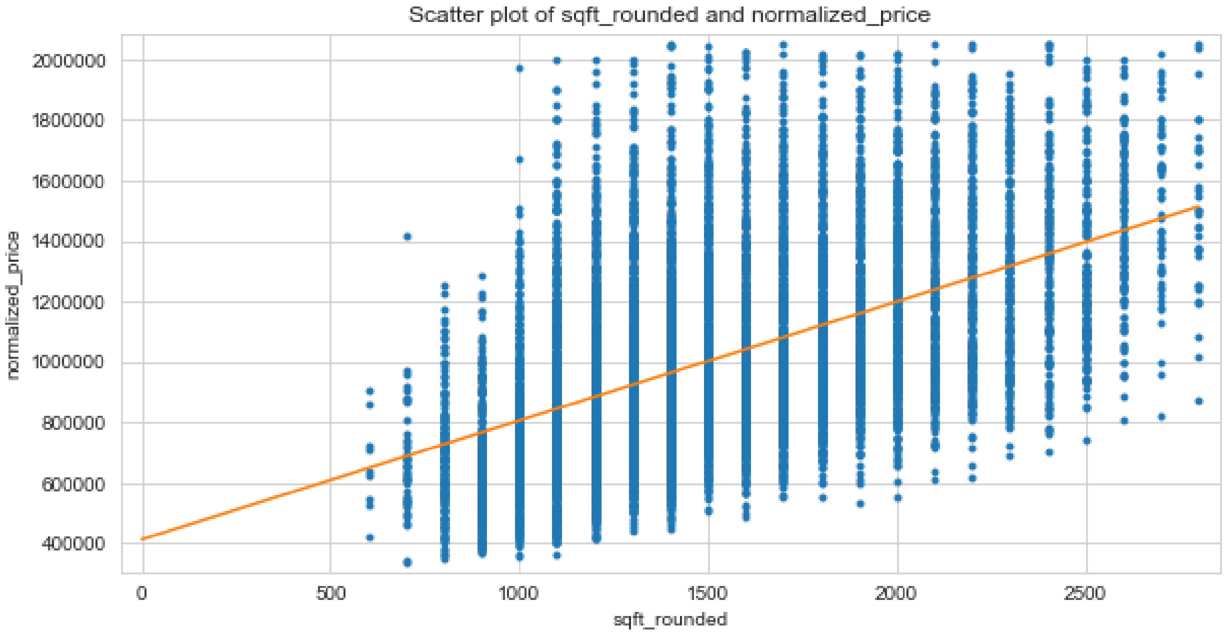
Scatter plot distribution



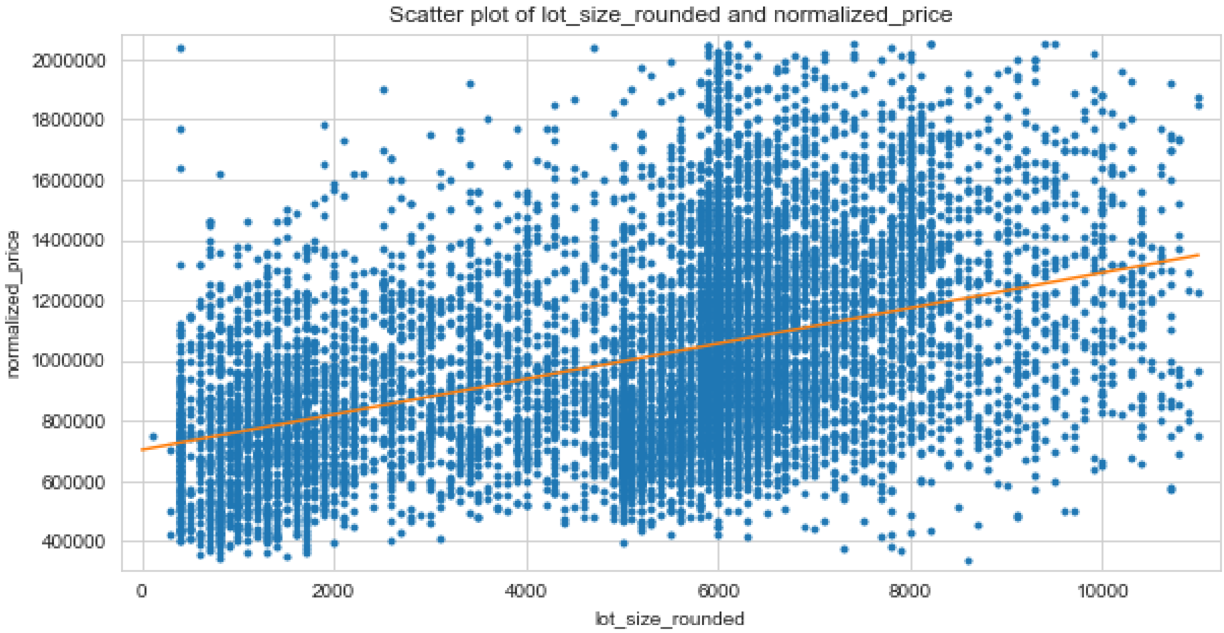
Number of bedrooms shows positive correlation with price



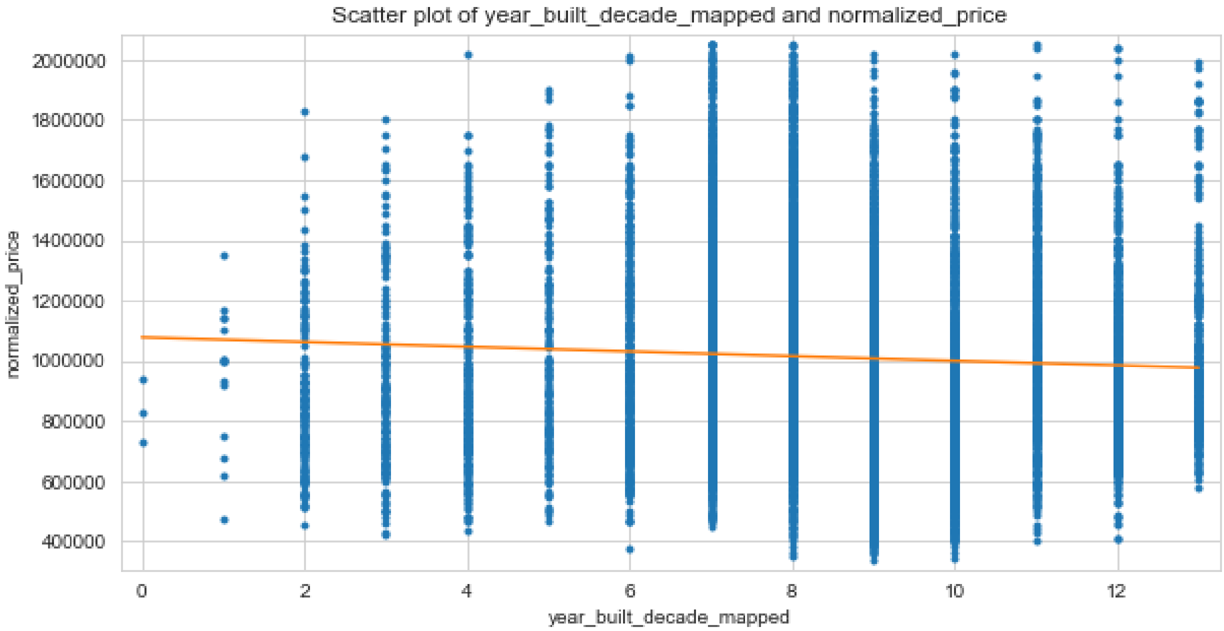
Number of bathrooms shows positive correlation with price



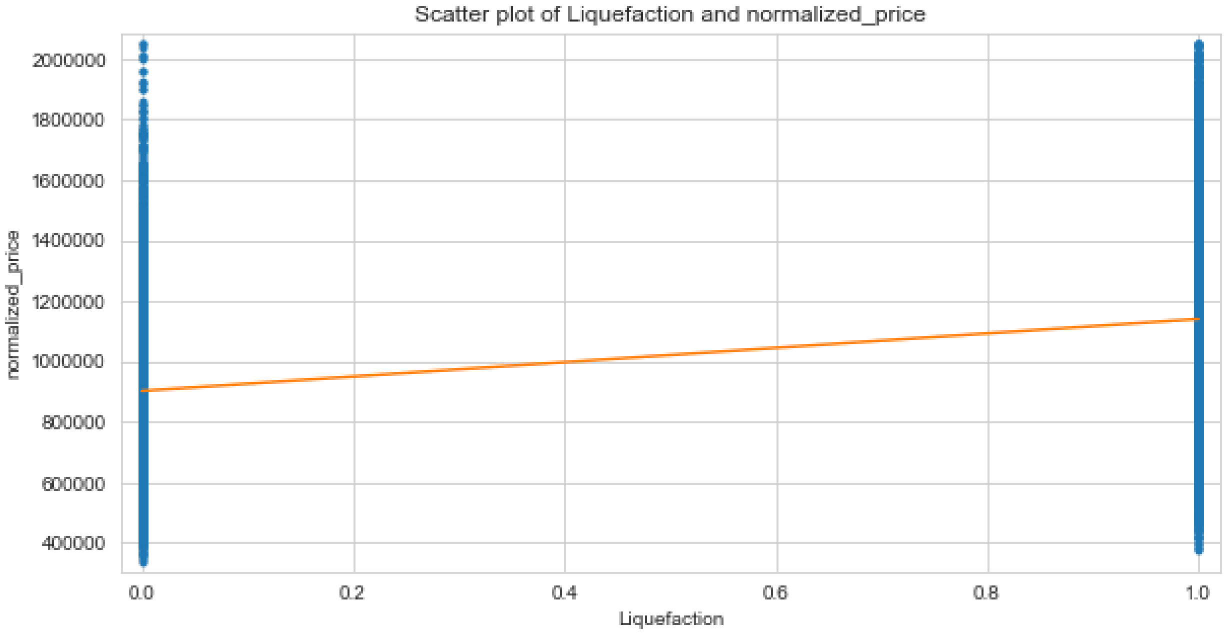
sqft shows positive correlation with price.



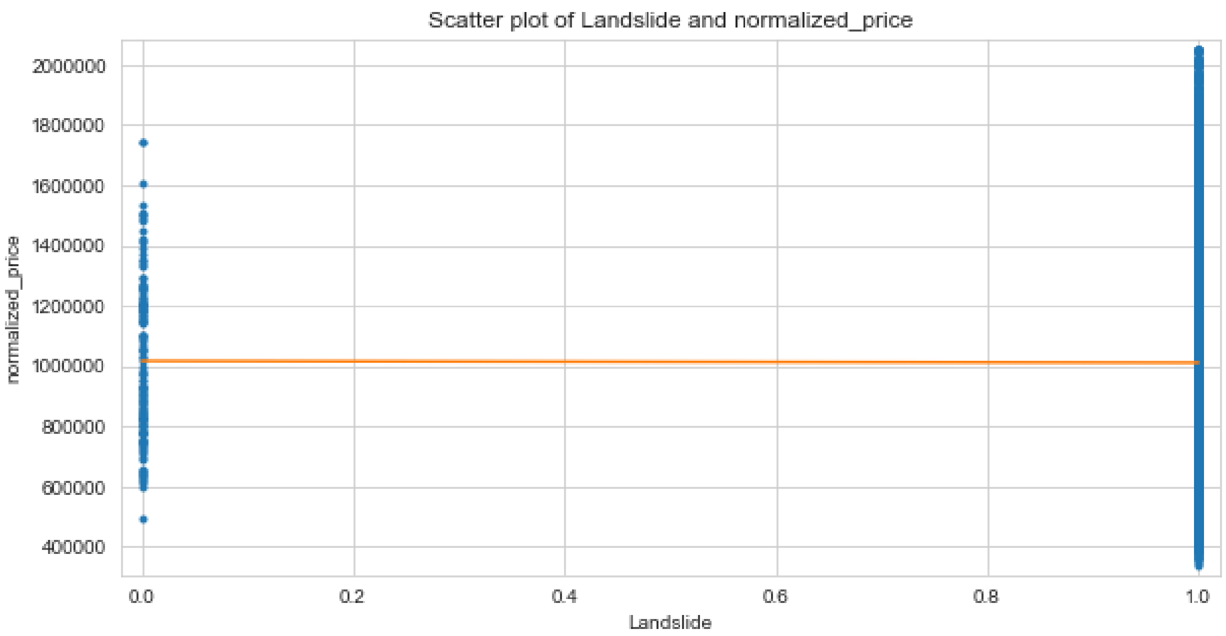
Lot size shows positive correlation with price



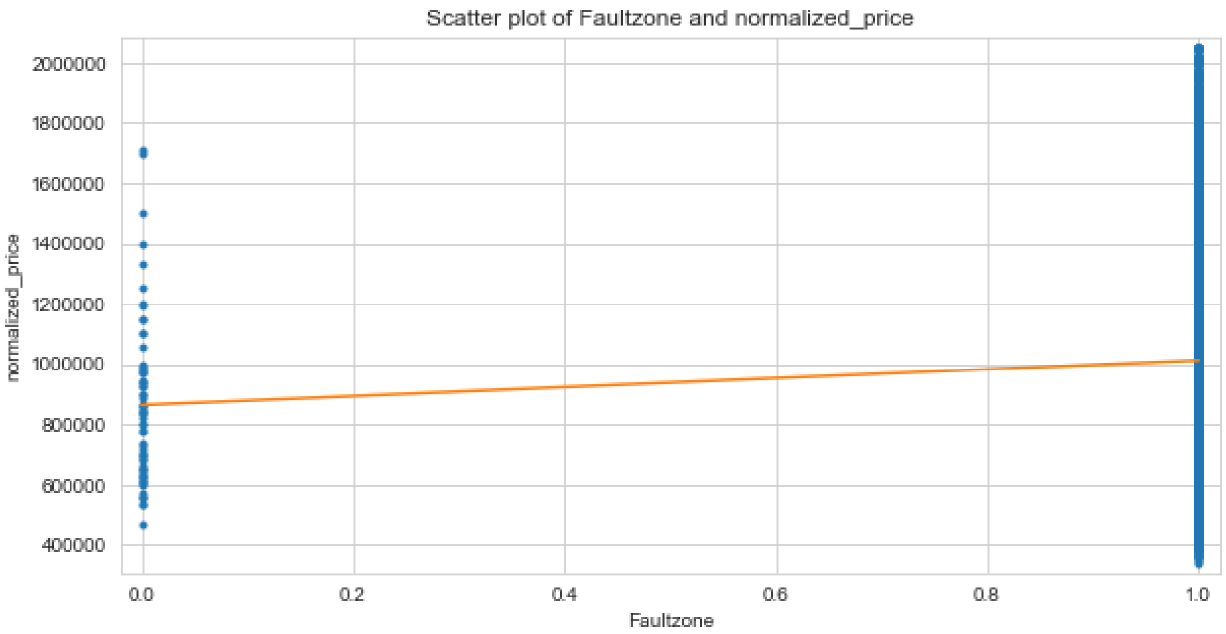
Year built shows weak negative correlation with price



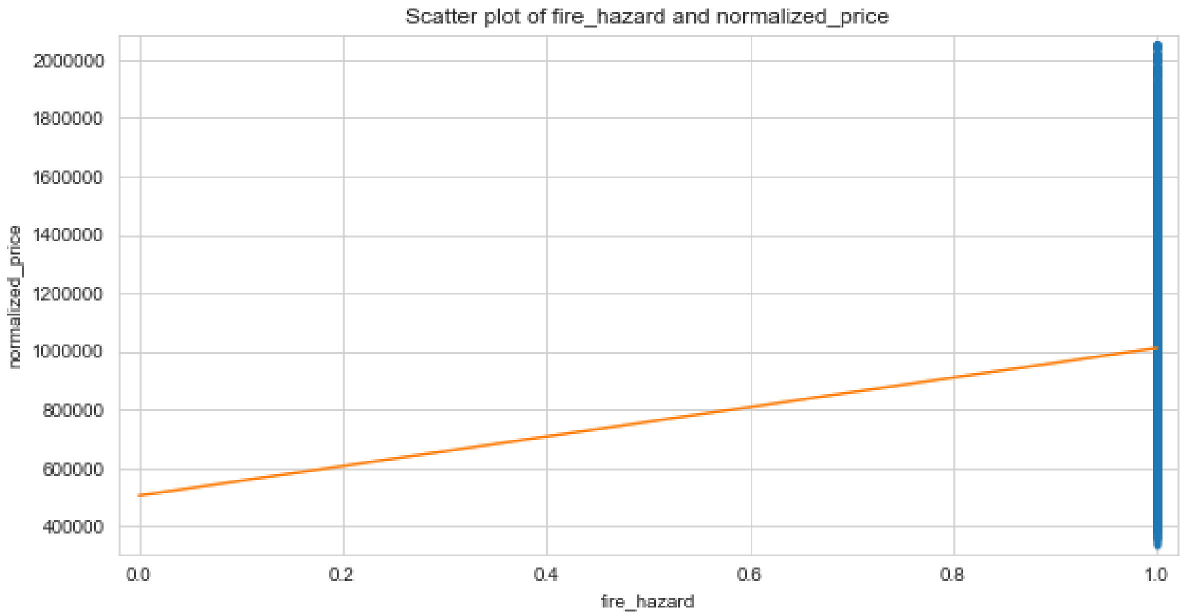
Liquefaction shows positive correlation with price



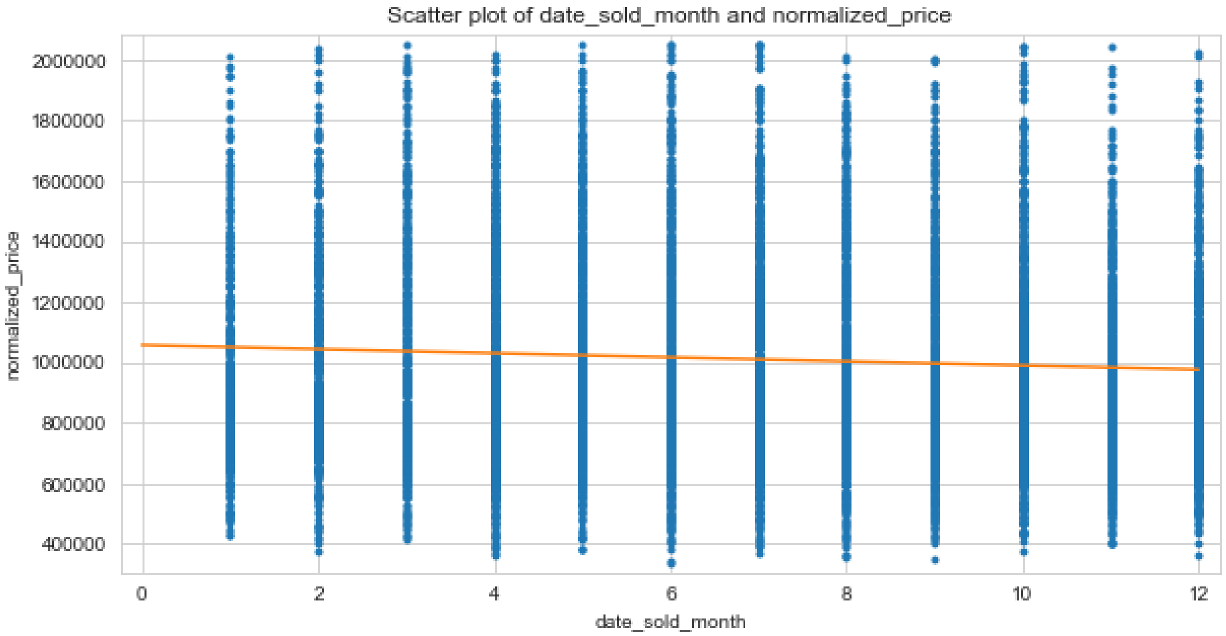
Landslide shows weak negative correlation with price



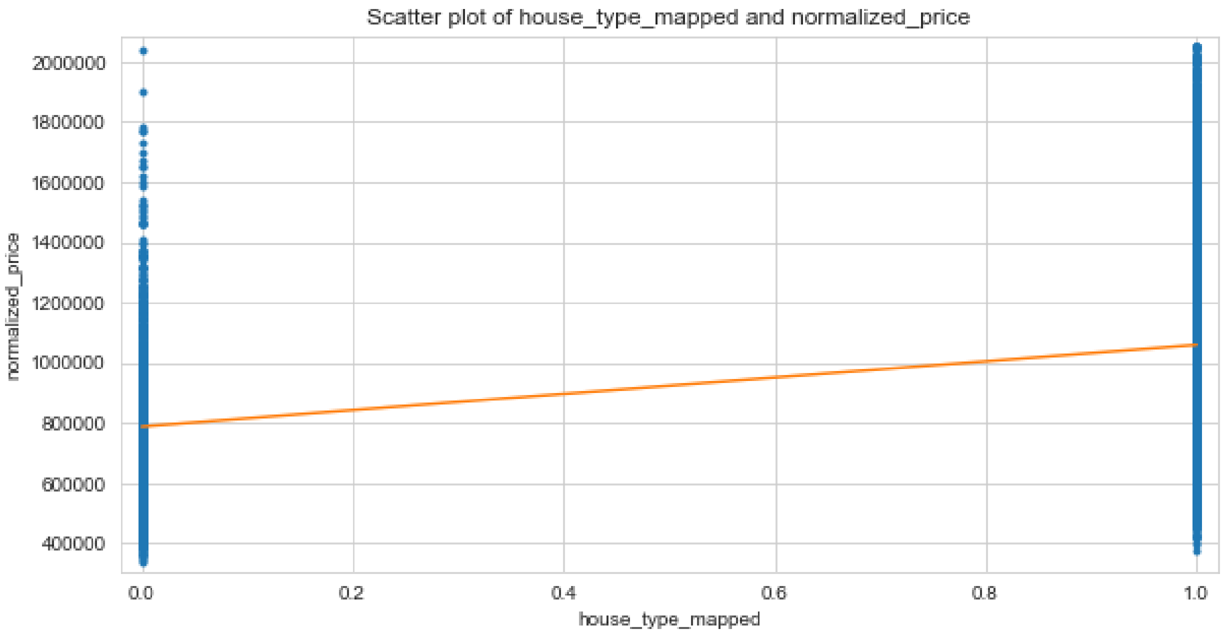
Fault zone shows positive correlation with price



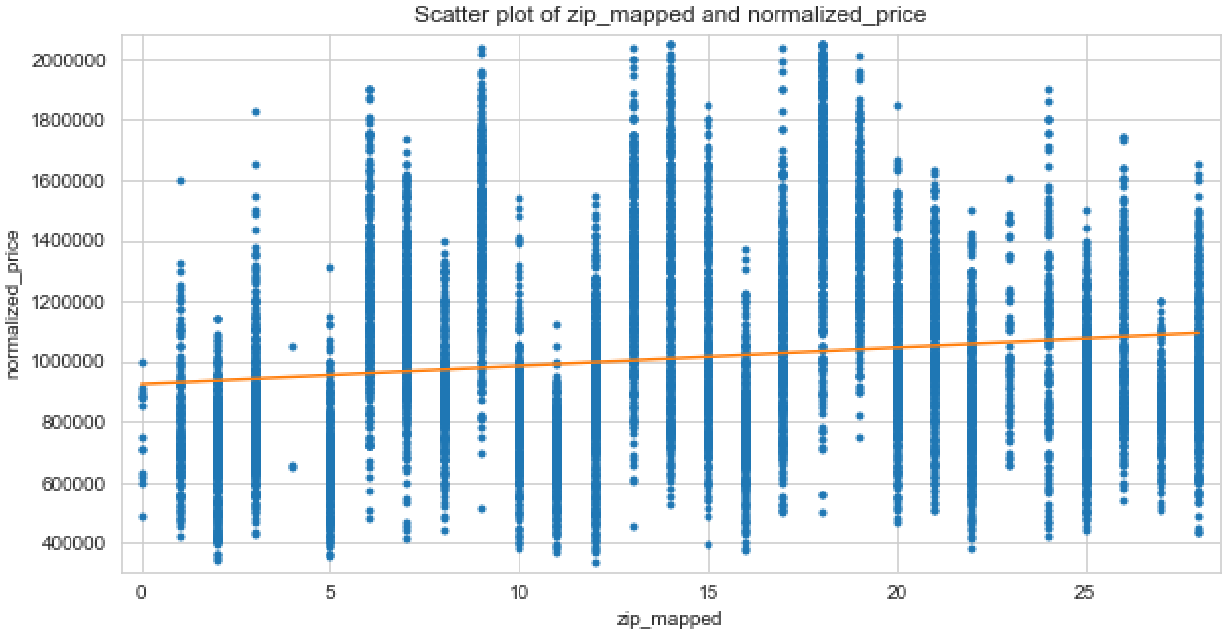
No correlation between fire hazard and price



Sold month shows negative correlation with price

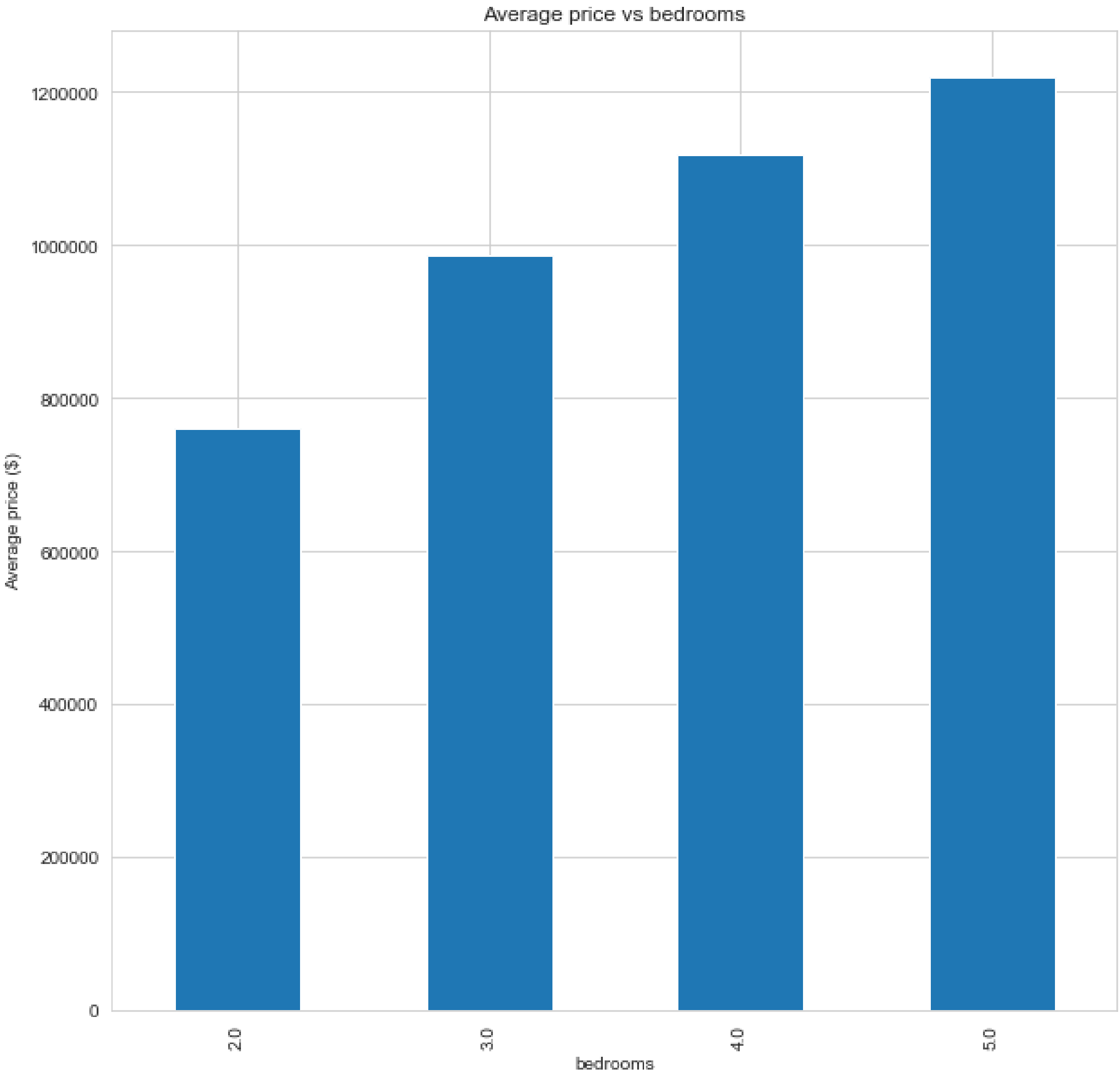


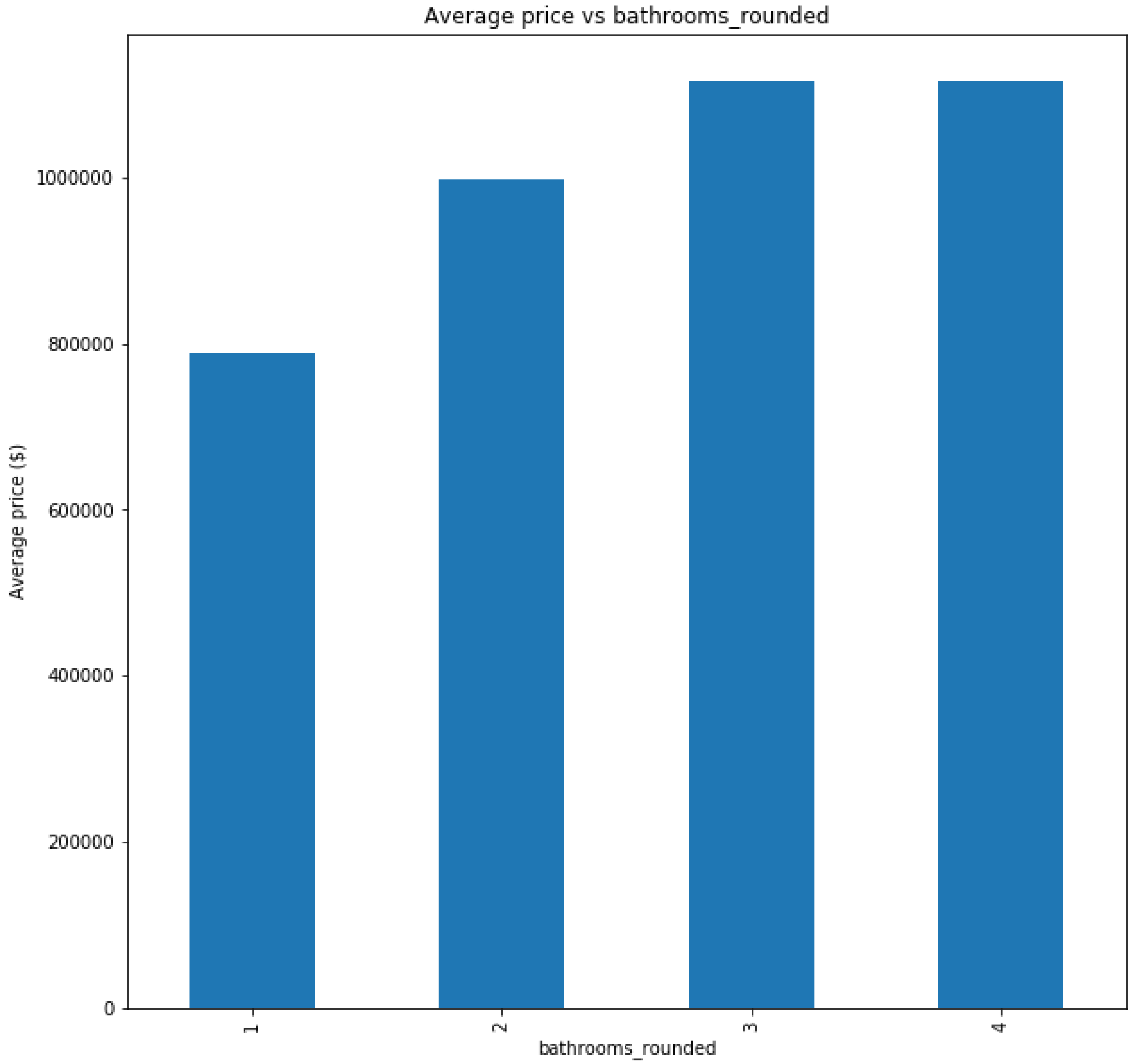
House type shows positive correlation with price

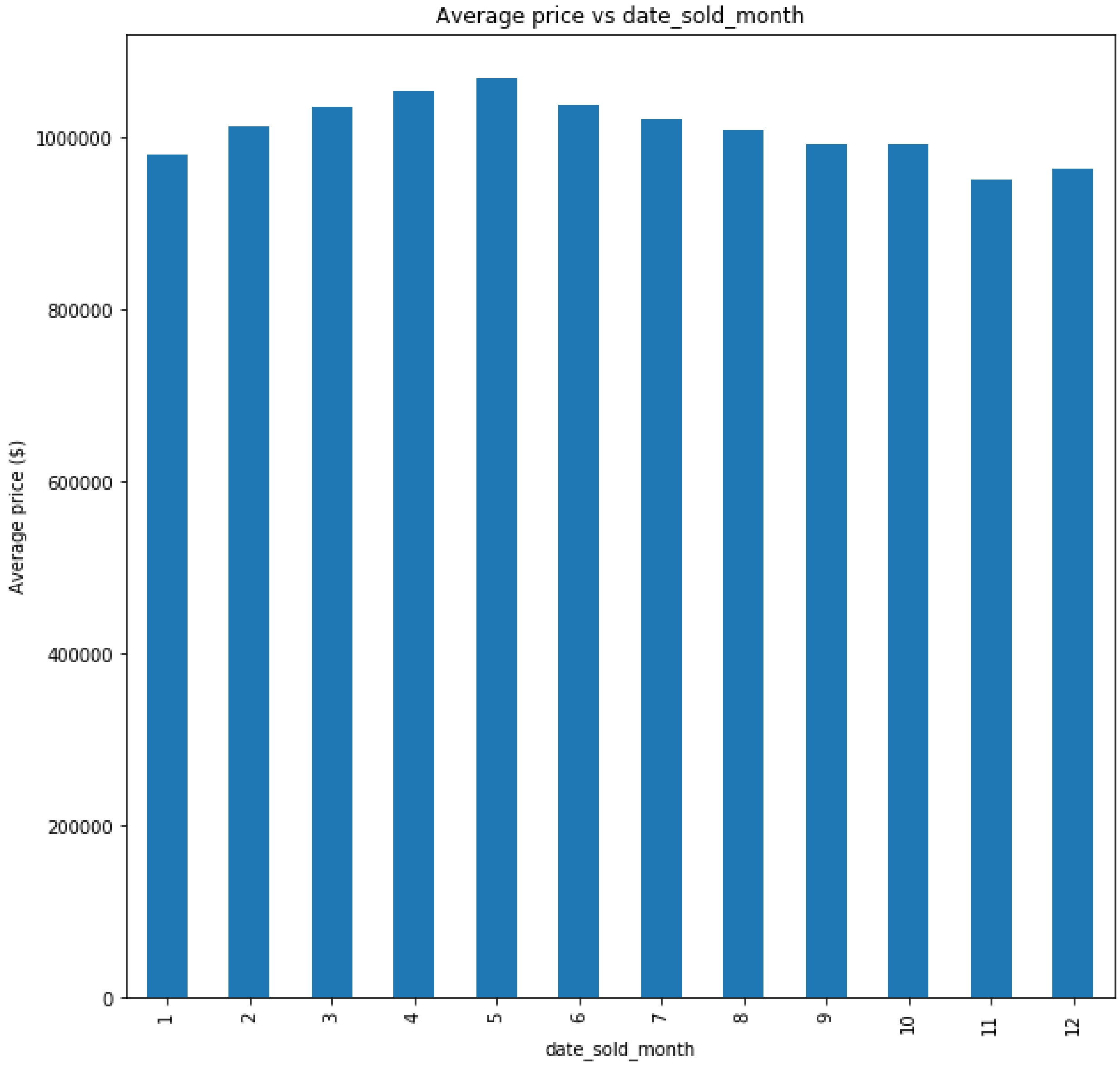


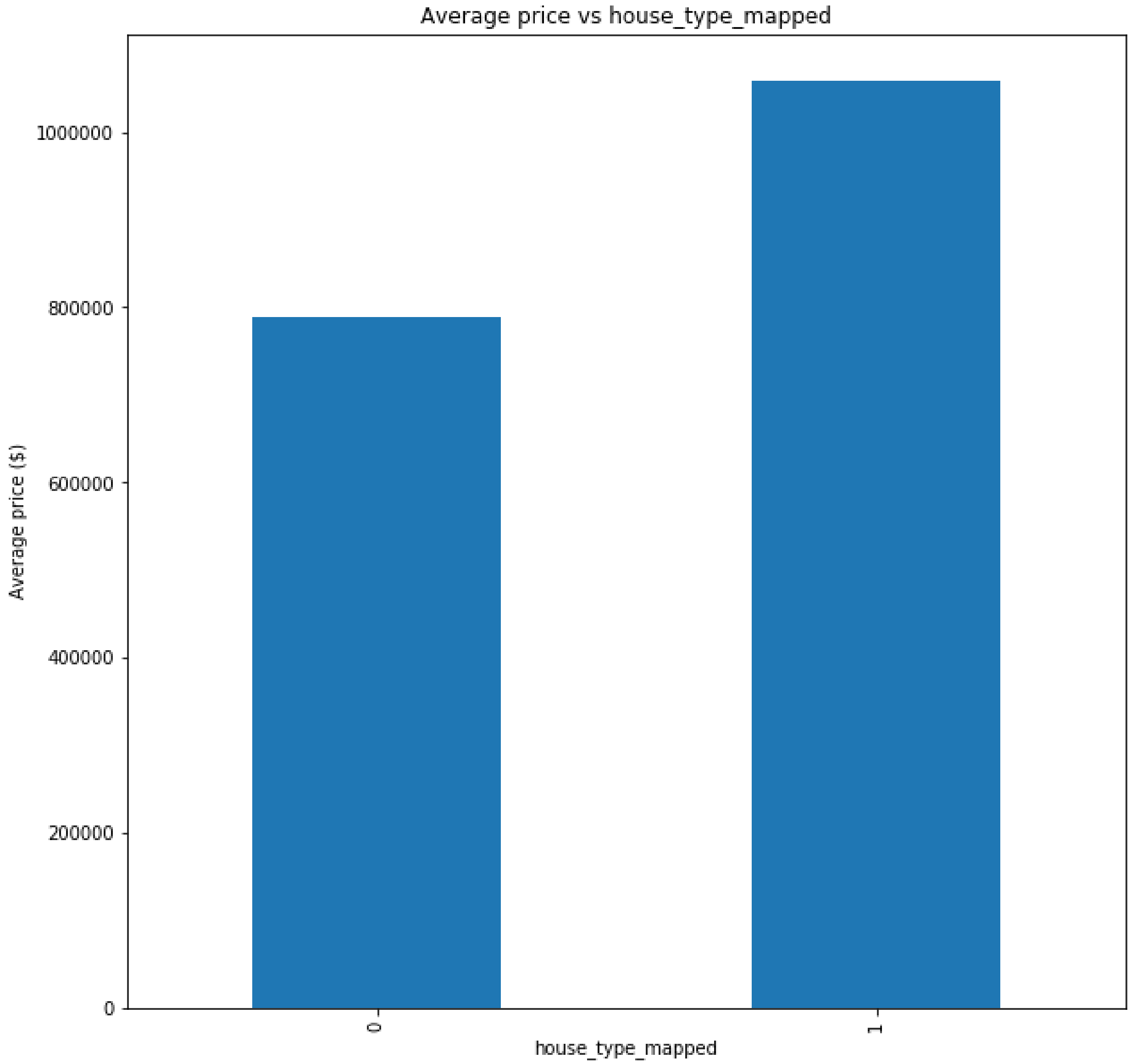
Zip code shows positive correlation with price

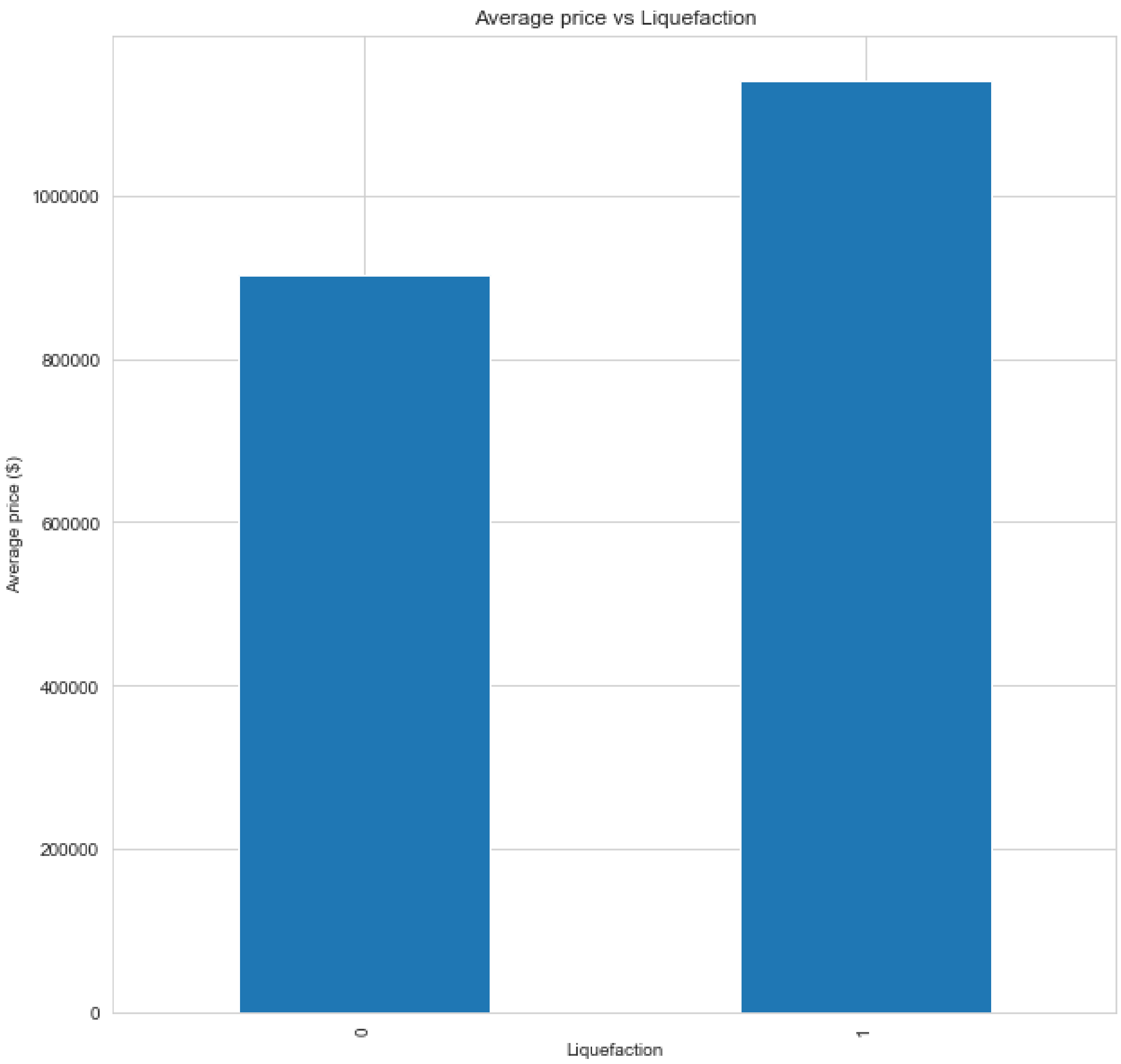
Average price bar plot distribution

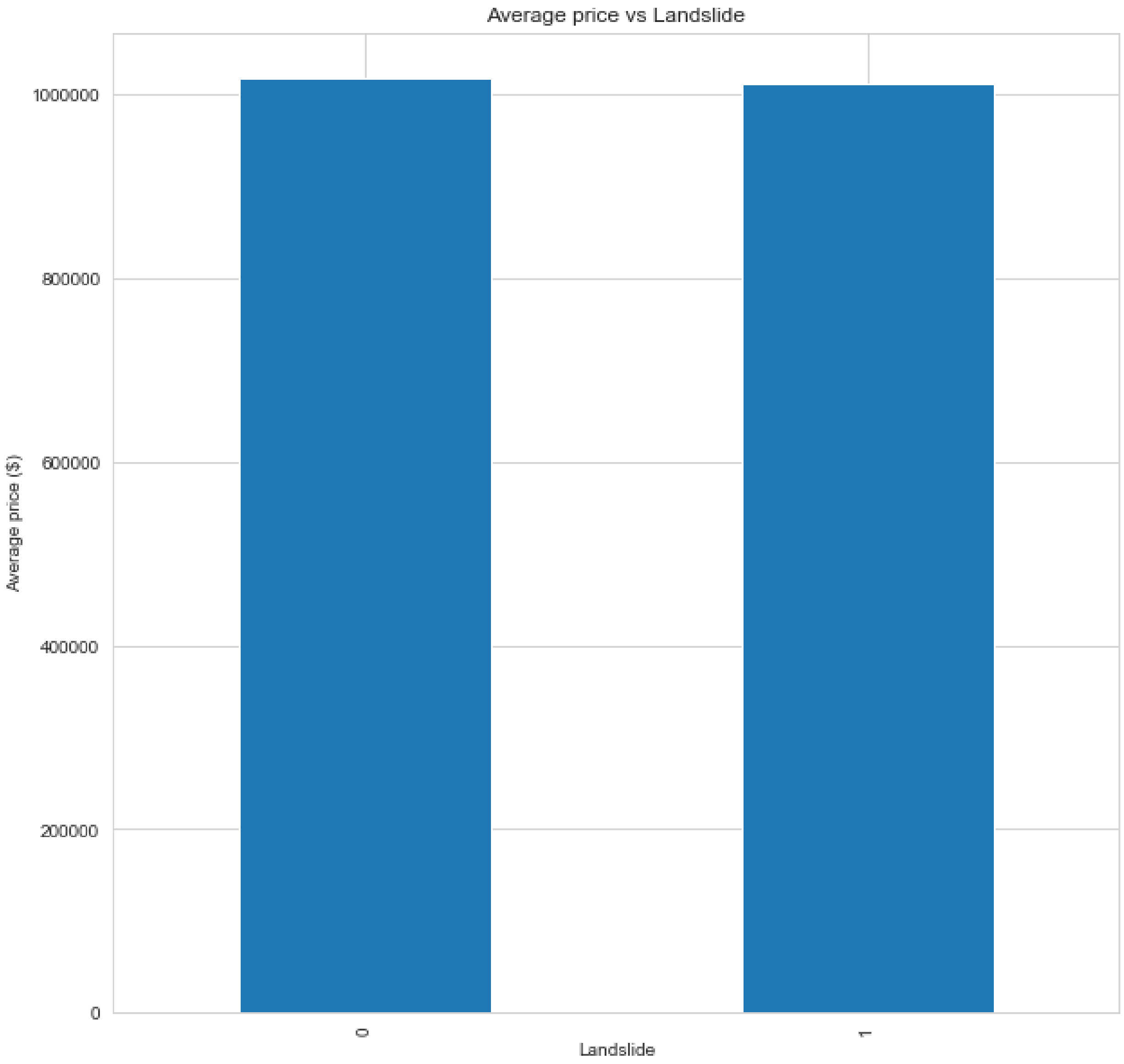


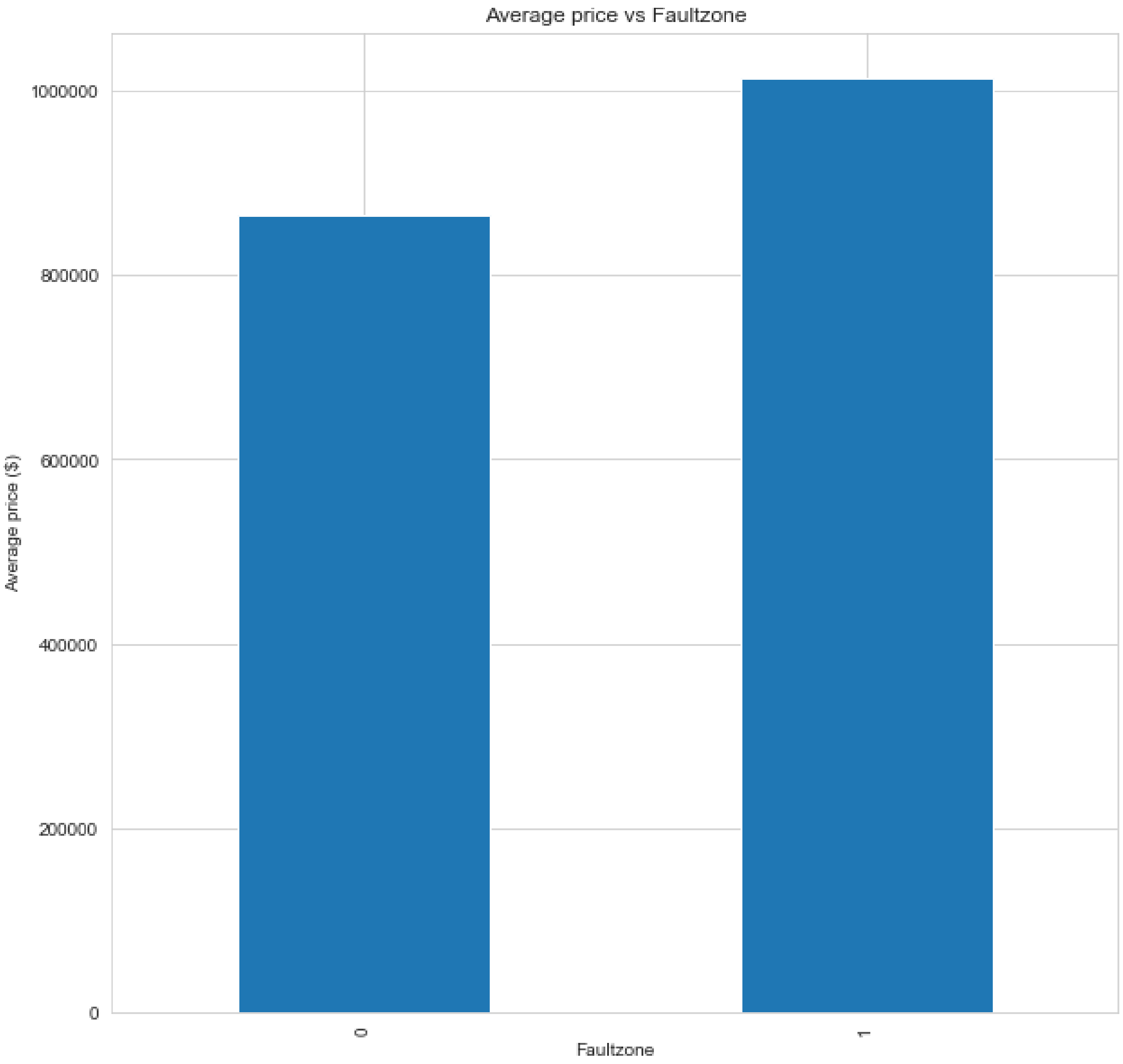

Average price vs bathrooms_rounded
Average price ($)
0
200000
400000
600000
800000
1000000
1
2
3
4
bathrooms_rounded


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


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


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


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


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

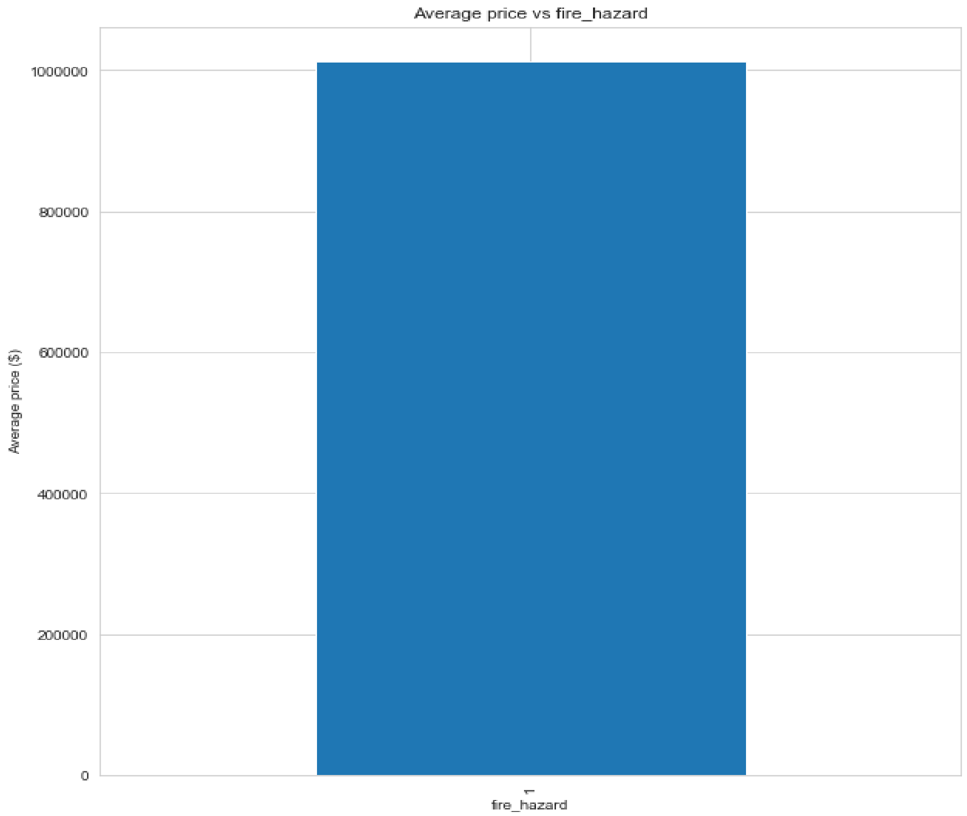


Fig 27: Average price distribution

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.

Single family EDA notebook:
https://nbviewer.jupyter.org/github/umaraju18/CapStone-Project/blob/master/sanjose_eda_single_family.ipynb
Townhouse EDA notebook:
https://nbviewer.jupyter.org/github/umaraju18/CapStone-Project/blob/master/sanjose_eda_townhouse.ipynb

MACHINE LEARNING MODELS

Introduction

In this project, different Regression models i.e. Linear Regression, Decision Tree Regression, Gradient Boosted Regression, and Random Forest Regression were used. The performance of those models using R^2 were compared. Based on these performance score, better performing model were suggested to predict house price.

First, the data was divided into independent variable X and dependent variable y. Independent variable X was used to predict the target variable y. The price, id and date column were dropped from the new_df dataframe to create the variable X. The price column from the new_df dataframe were used to create the variable y. In this project, different metrics were used to the performance of the Regression models such as Mean squared errors, Root mean squared errors, R-squared score, Mean absolute deviation, Mean absolute percent errors, etc. In this project, Root mean squared error and R-squared score were used to evaluate the performance of the regression model. In order to save the metrics of the model, a data frame was created and it was named metrics. Next, the data was splitted into training and testing set. 80% of the randomly selected data were kept as a training set and 20% of the randomly selected data as a testing set. The model was learned using the 80% of the data, and the rest 20% testing data were used as an unseen future dataset to predict the house price.

Linear Regression Model was built using the default parameters, and the model was fitted using the training dataset. X_test data was used to predict using the model. Then, Mean squared error (MSE), Root mean squared error (RMSE), R-squared score (r2_score), Mean absolute deviation (MAD), and Mean absolute percent error (MAPE) were calculated.

```
Mean Squared Error (MSE): 54296634556.66
Root Mean Squared Error (RMSE): 233016.3826
r2_score: 0.4637
Mean Absolute Error (MAE): 182963.47
Mean Absolute Percent Error (MAPE): 19.28
```

Feature Selection

Backward elimination method of feature selection were used. Feature selection is the process of selecting a subset of relevant features that may improve the performance of the model. First, the worst attribute from the feature was removed. The date_sold_month were removed because it has a very weak correlation with the price of the house. Then, year_built_decade_mapped were removed from the feature set. Then, a univariate feature selection package called SelectKbest from the sklearn library was tried. Below are correlation coefficient for different features.

date_sold_month	-0.065022
Year_built_decade	-0.052772
Landslide	-0.002046
Faultzone	0.036062
zip_mapped	0.139833
bathrooms_rounded	0.250558
house_type_mapped	0.317384
Liquefaction	0.361638
bedrooms	0.368206
lot_size_rounded	0.428212
sqft_rounded	0.521310
fire_hazard	NaN

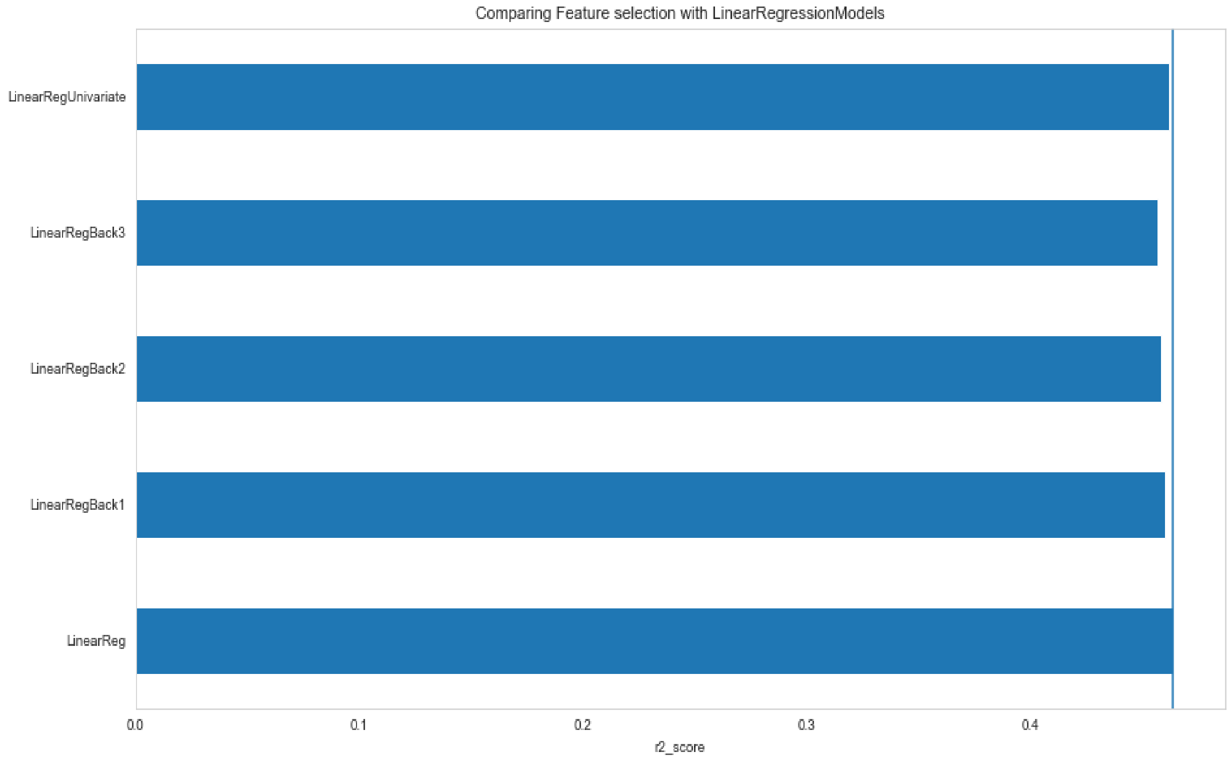


Fig 1a: Comparing Feature selection with Linear Regression models using R2 score

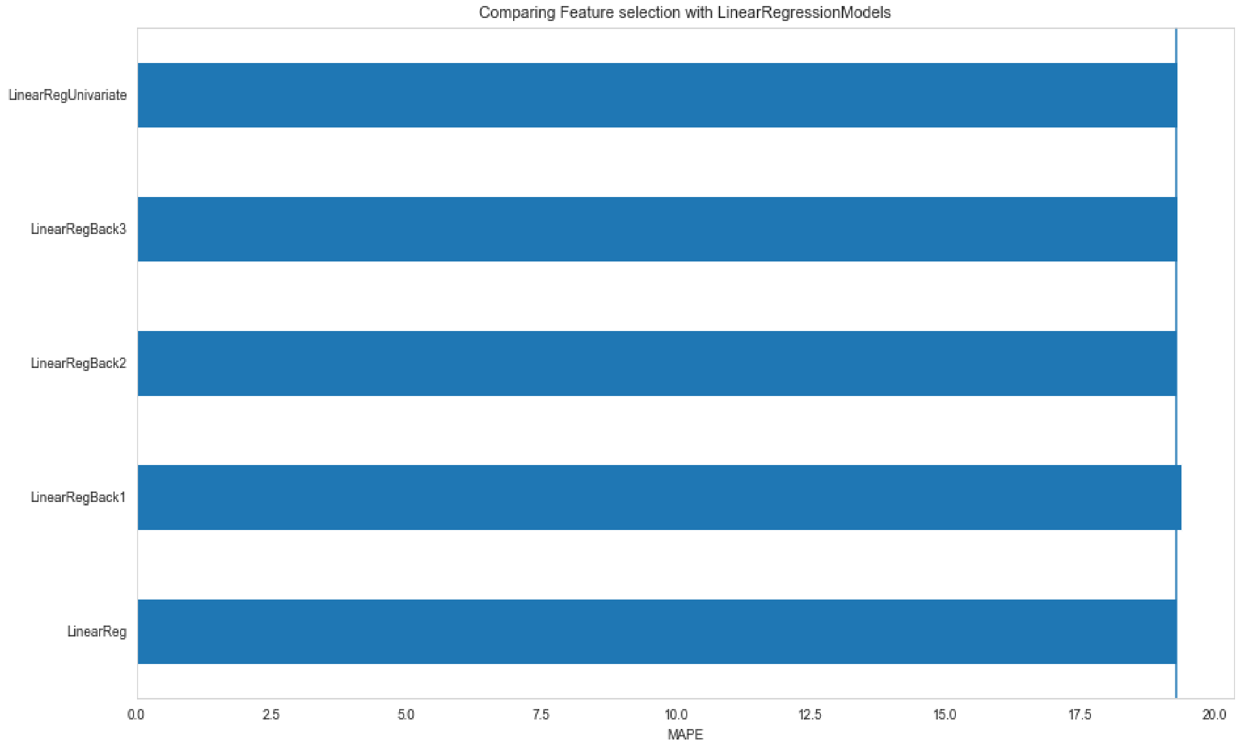


Fig 1b: Comparing Feature selection with Linear Regression models using MAPE score

Comparing all different Linear Regression model with feature selection, both bar plots (Fig 1a and 1b) suggest that we should keep all the features to better predict the house price.

Decision Tree Regression

Decision Tree Regressor Model using the default parameters were built and listed performance score below. The sample decision Tree regression flowchart is shown in Fig 2

Mean Squared Error: 43361291535.89
Root Mean Squared Error: 208233.74
r-squared score : 0.5716779806579388
Mean Absolute Deviation (MAE): 155321.44
Mean Absolute Percent Error (MAPE): 16.13

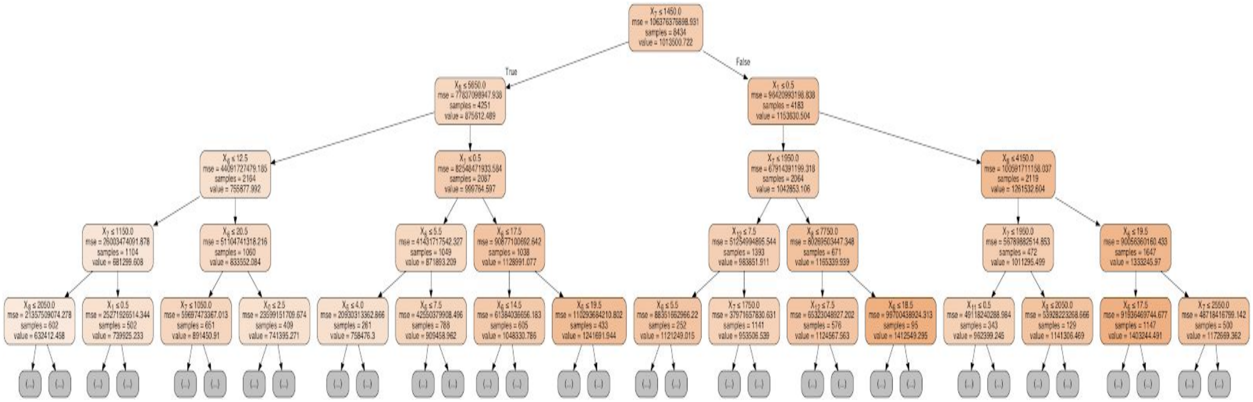


Fig 2. Sample Decision Tree Diagram

Gradient Boosting Regression

Gradient regression Model using the default parameters were built and listed performance score below.
Mean Squared Error: 25679823410.25
Root Mean Squared Error: 160249.25
r-squared score : 0.7463351890632407
Mean Absolute Deviation (MAE): 124156.64
Mean Absolute Percent Error (MAPE): 13.02

Random Forest regression

Random Forest regression Model using the default parameters were built and listed performance score below.
Mean Squared Error (MSE): 25762504105.35
Root Mean Squared Error (RMSE): 160507.02
r-squared score : 0.745518470717669
Mean Absolute Deviation (MAE): 122098.99
Mean Absolute Percent Error (MAPE): 12.92

The horizontal bar plot (Fig 3) is shown below to compare r2_score and MAPE for different regressor.

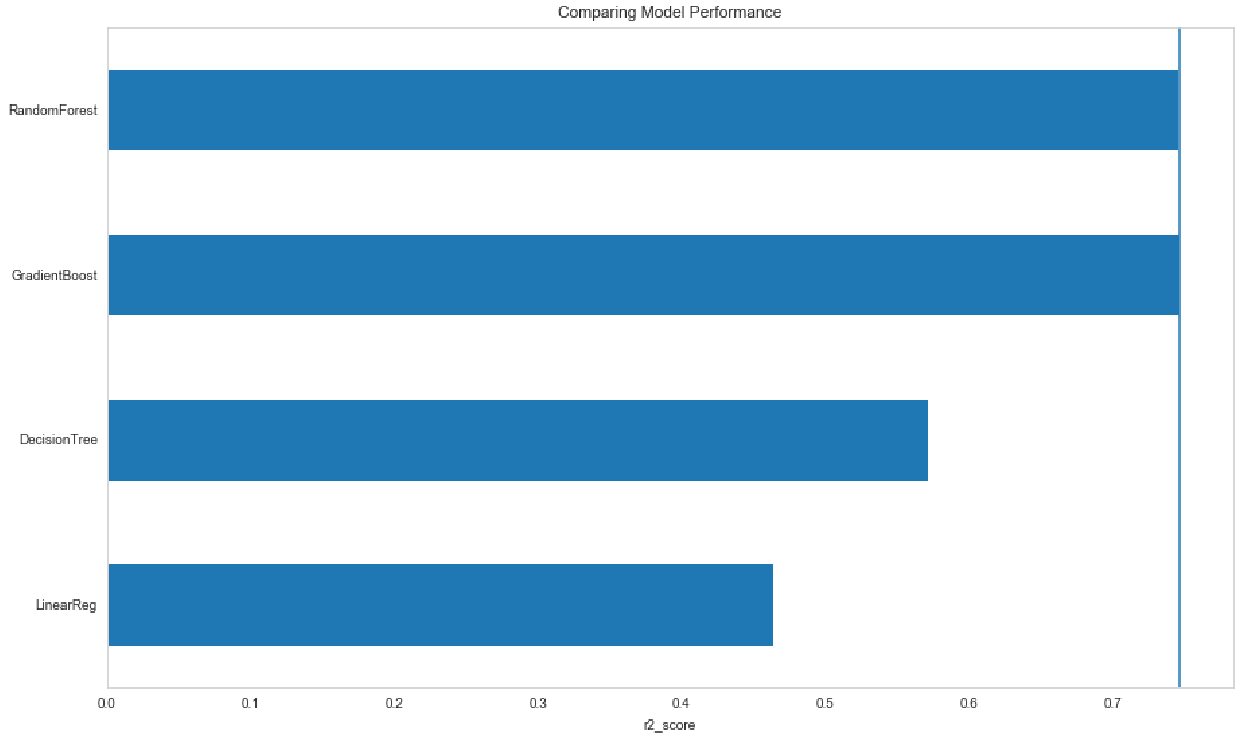


Fig 3: Comparing model performance for different regression models using r2 score

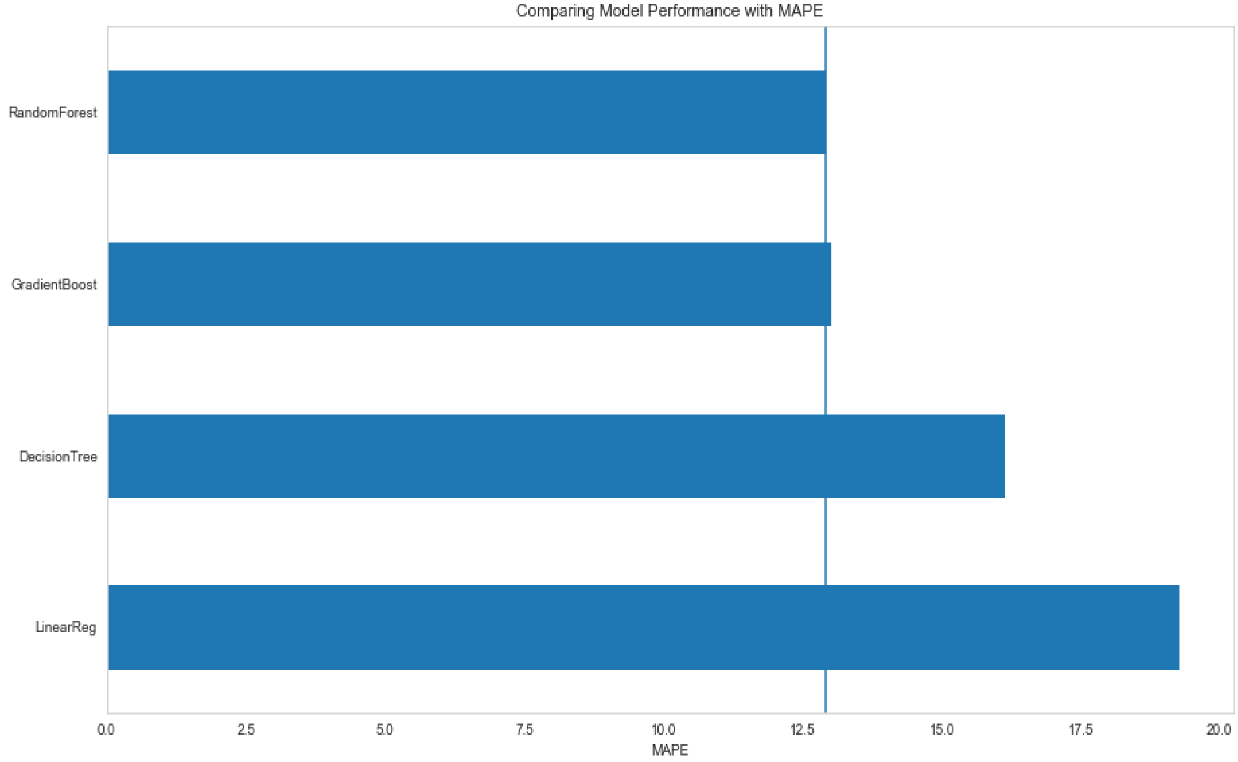


Fig 4. Comparing model performance for different regression models using MAPE score

According to r2_score (Fig 3), Gradient Boosted Regression model was the best performing model. The random Forest Regression model was the second better performing model for this dataset.

According to Mean absolute percentage error (Mape) (Fig 4), Random Forest Regressor model was the better performing model. Gradient Boosting Regressor model was the second better performing model.

Hyperparameter Tuning

Next, GridSearchCV were used to tune the Hyperparameters of the model to improve the performance of the model. GridSearchCV is a cross-validation method which allows us to use a set of parameters that we want to try with a given model. Each of those parameters was used to perform cross-validation and finally, the best parameters for the model is saved. A new dataframe was created and was named tuned_metrics to save the metrics of the tuned models. The method .get_params() was used to find all the parameters of the model. For Linear Regression model, 'copy_X', 'fit_intercept', and 'normalize' parameters were used inside the param_grid. The param_grid is a dictionary with parameters names as keys and lists of parameter settings to try as values. cv = 5 was used, which is the number of folds used. We can get the best parameter for any Regression model for this data set using the best_params_ attribute of GridSearchCV. The Linear Regression model and Decision Tree Regressor model were tuned using GridSearchCV. Then, the Gradient Boosting Regressor model and Random Forest Regressor model were tuned using RandomizedSearchCV. RandomizedSearchCV helped to minimize the computation time because GridSearchCV can take very long computational time to for both Gradient Boosting Regressor and Random Forest Regressor. Tuned model performance comparison for different regression models using RMSE score is shown in Fig 5.

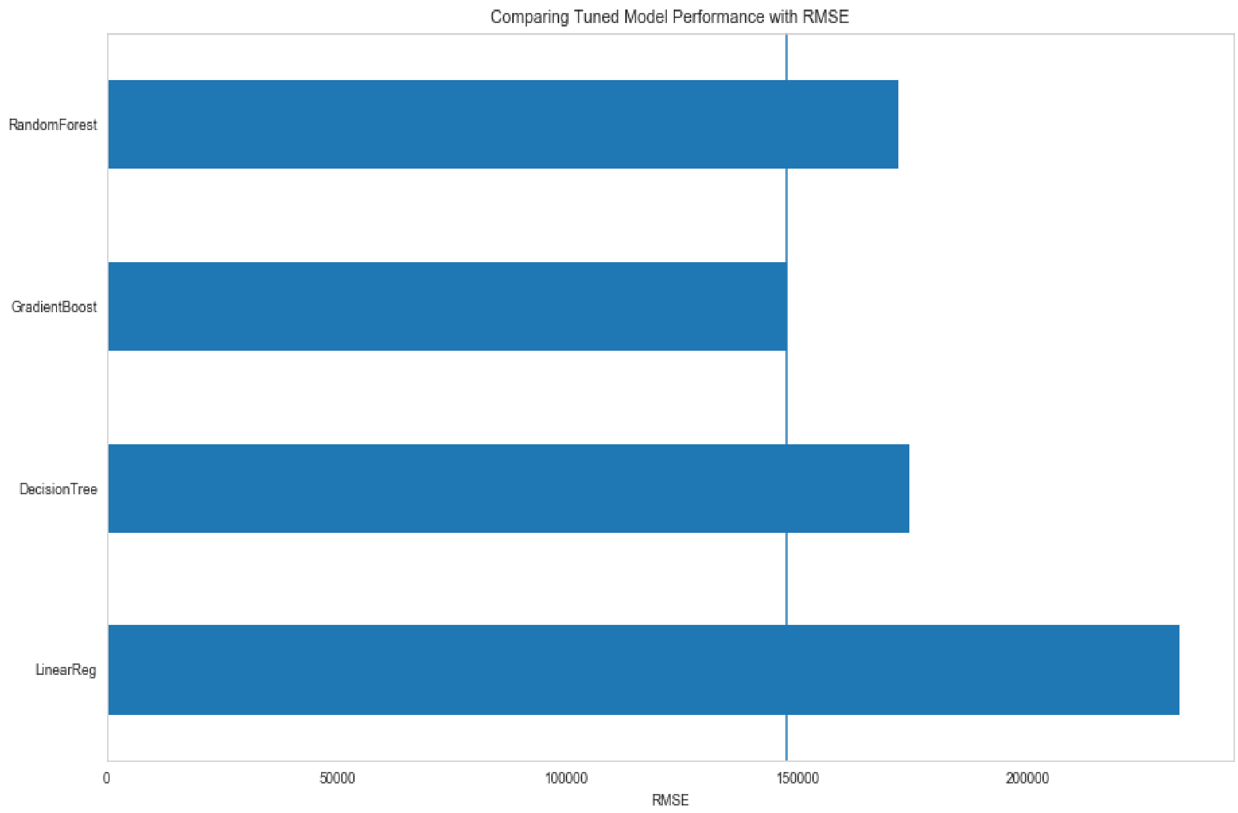


Fig 5. Comparing tuned model performance for different regression models using RMSE score

According to Root Mean Squared Error (Fig 5), Gradient Boost Regression model was the best performing model with the lowest error 147773.

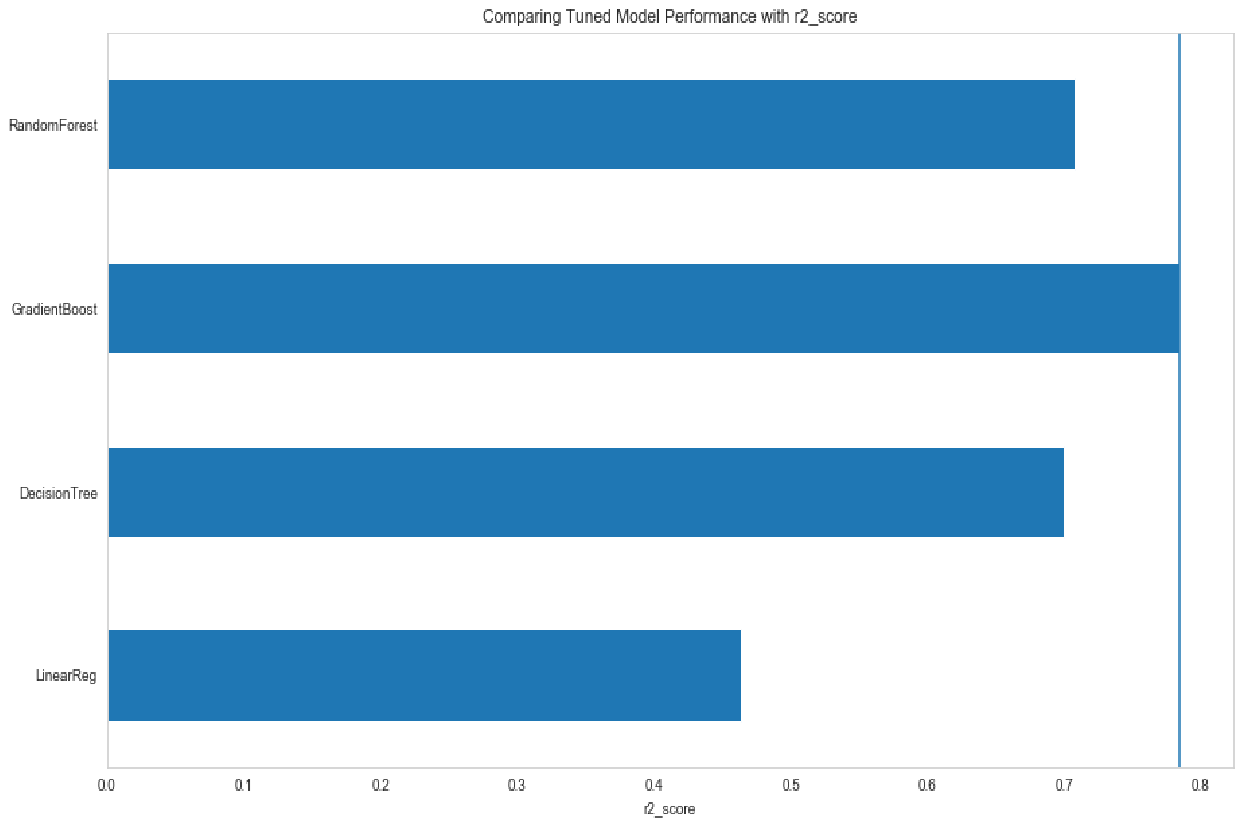


Fig 6. Comparing tuned model performance for different regression models using R2 score

According to r2_score (Fig 6), Gradient Boost Regression model was the best performing model with the highest score of 0.78.

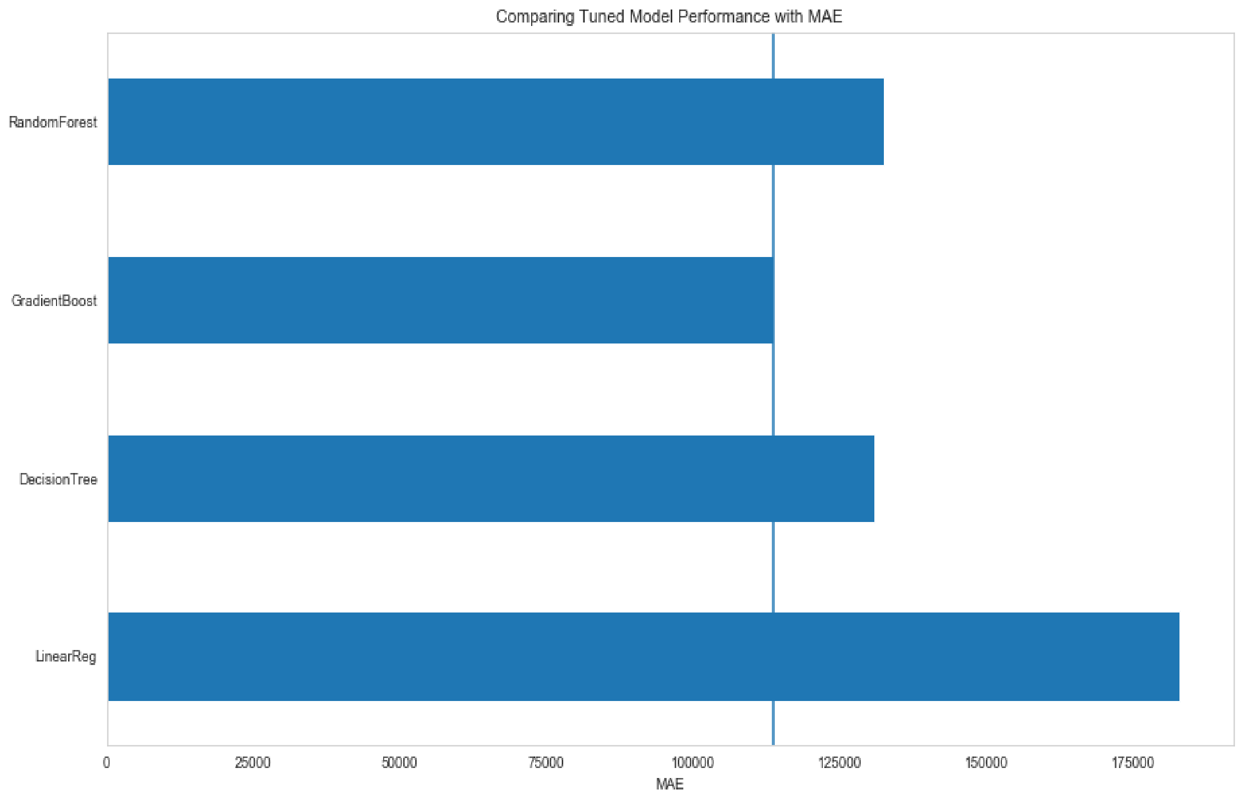


Fig 7. Comparing tuned model performance for different regression models using MAE score

According to Mean Absolute Error (MAE) (Fig 7), Gradient Boost Regression model was the best performing model with lowest error 113737.

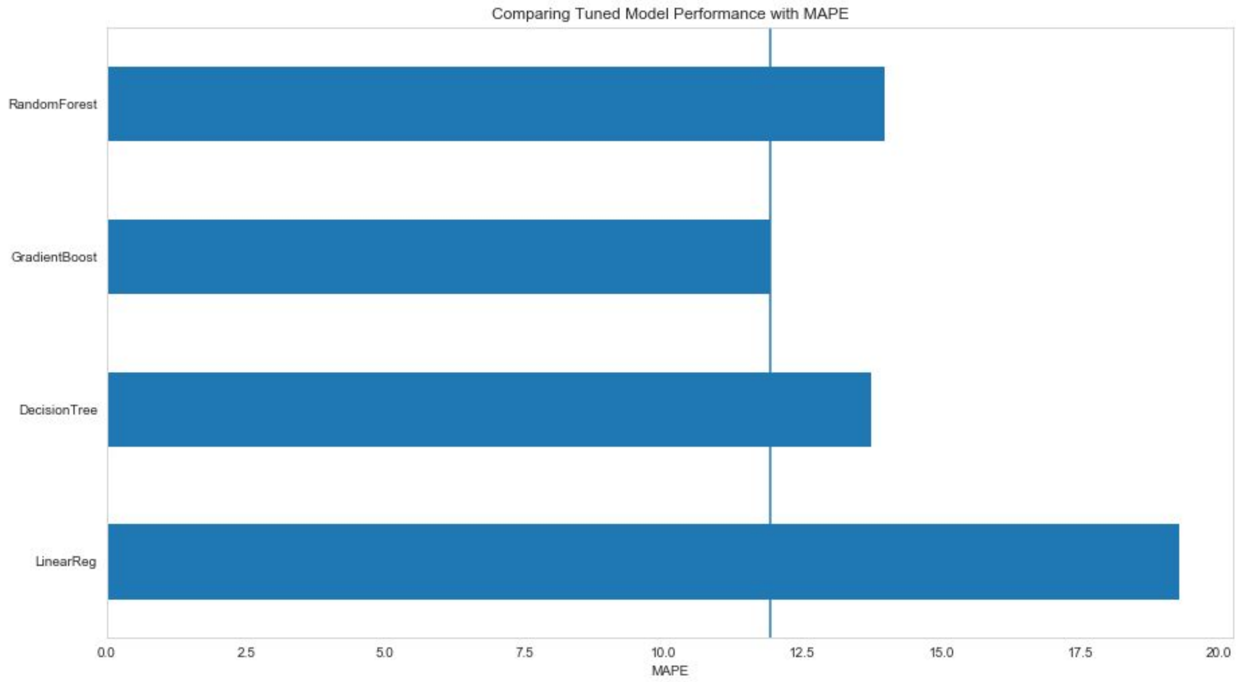


Fig 8. Comparing tuned model performance for different regression models using MAPE score

According to the Mean absolute percentage error (MAPE) (Fig 8), Gradient Boost Regression model was the best performing model with lowest percentage error 12%. All of the metrics suggest that Gradient Boosted Regression model was the better performing model for this dataset.

Summary

Gradient Boosting Regression model was a good model to predict house price because it was better than a random guess and it outperformed the other three Regression Model. The model may be improved in the future with more data collection. Many other Regression models which were not included in this project can also be built and tried. For future work, I would recommend this Gradient Boosting Regression model to predict house price.

Ipython notebook for machine learning:
https://nbviewer.jupyter.org/github/umaraju18/CapStone-Project/blob/master/sanjose_machine_learning_model.ipynb

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