House Price Prediction in Natural Hazard Prone Areas

Capstone Project 1: Data Wrangling

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.

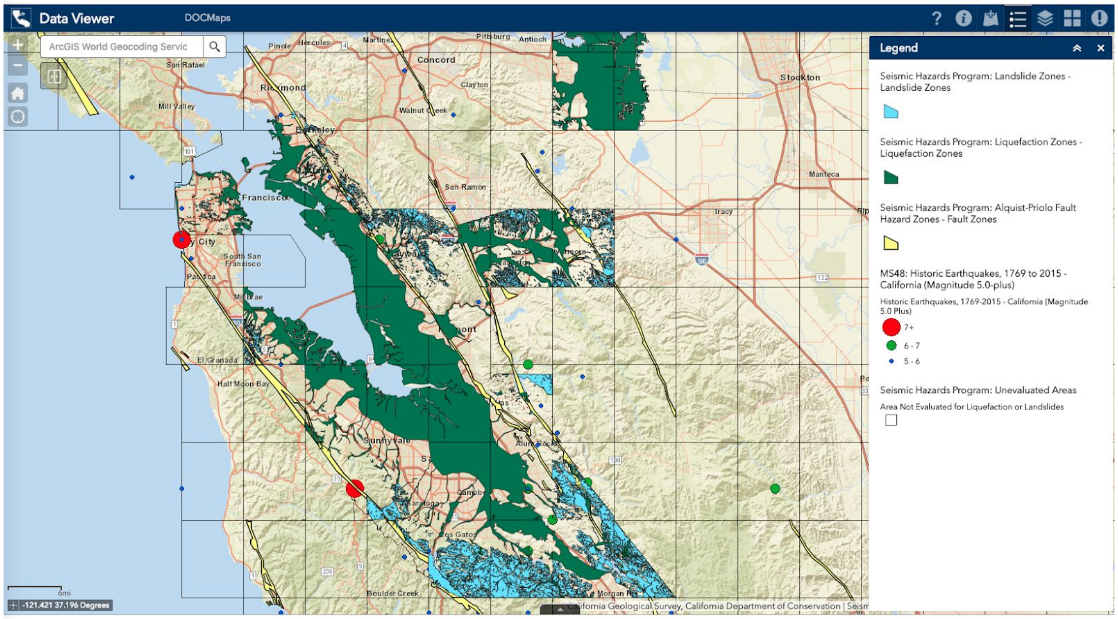

Data Viewer
DOCMaps
ArcGIS World Geocoding Servic
Legend
Seismic Hazards Program: Landslide Zones - Landslide Zones
Seismic Hazards Program: Liquefaction Zones - Liquefaction Zones
Seismic Hazards Program: Alquist-Priolo Fault Hazard Zones - Fault Zones
MS48: Historic Earthquakes, 1769 to 2015 - California (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
Pinole
Hercules
Martinez
Pittsburg
Antioch
Concord
Clayton
Richmond
Walnut Creek
Berkeley
Stockton
Manteca
San Ramon
Tracy
Francisco
South San Francisco
Pacifica
Millbrae
El Granada
Half Moon Bay
Sunnyvale
Campbell
Los Gatos
Boulder Creek
Morgan Hill
6mi
-121.421 37.196 Degrees
California Geological Survey, California Department of Conservation | Seism

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

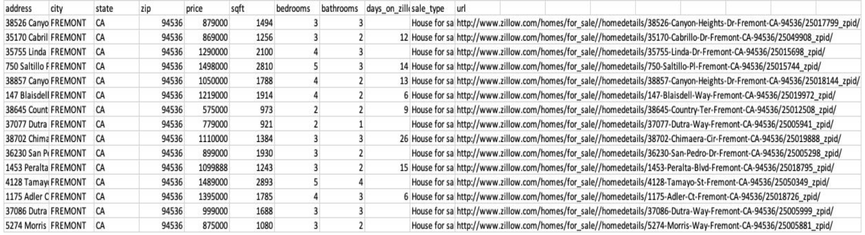
3. *Wildfire hazard potential data:*
   http://www.fire.ca.gov/fire_prevention/fhsz_maps/FHSZ/santa_clara/San_Jose.pdf

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)

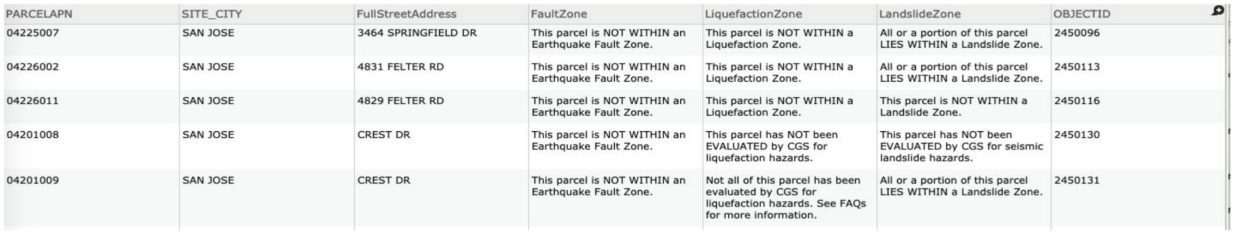


address	city	state	zip	price	sqft	bedrooms	bathrooms	days_on_zill	sale_type	url
38526 Canyo	FREMONT	CA	94536	879000	1494	3	3		House for sa	http://www.zillow.com/homes/for_sale//homedetails/38526-Canyon-Heights-Dr-Fremont-CA-94536/25017799_zpid/
35170 Cabril	FREMONT	CA	94536	869000	1256	3	2	12	House for 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
```

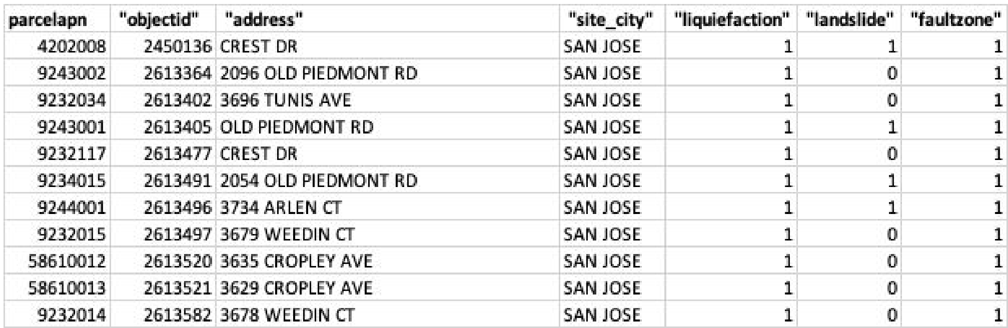
If landslide Zone:

```
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.

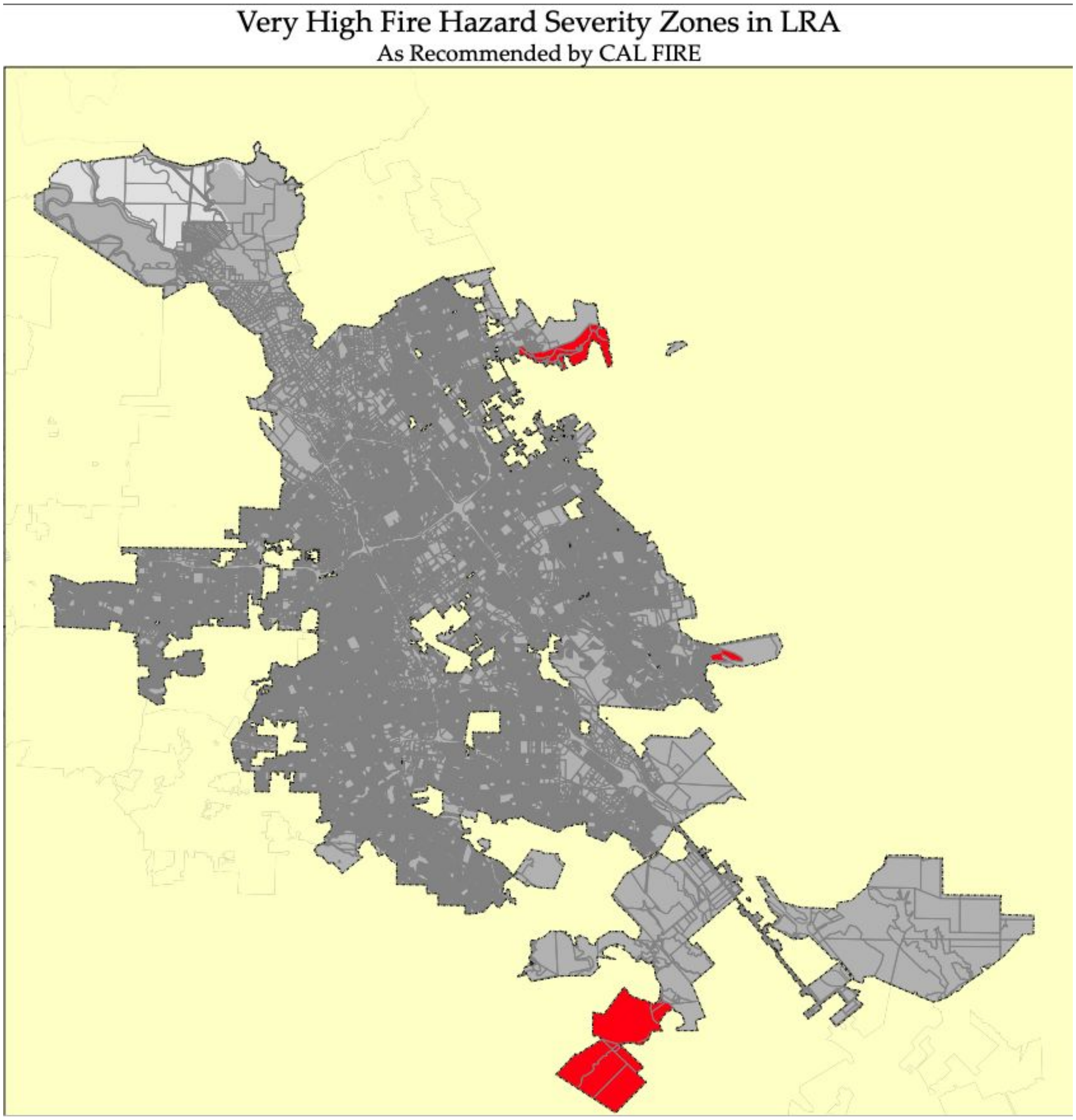

Very High Fire Hazard Severity Zones in LRA
As Recommended by CAL FIRE

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

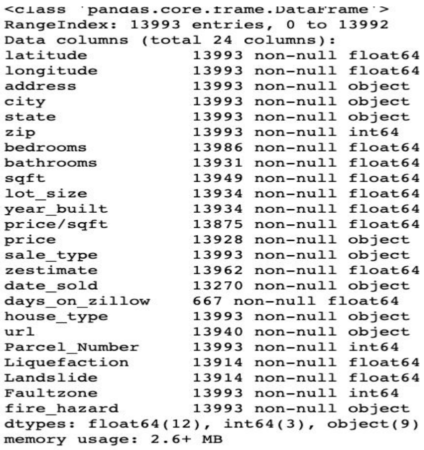
Fig 5b: Sample fire hazard data

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
```

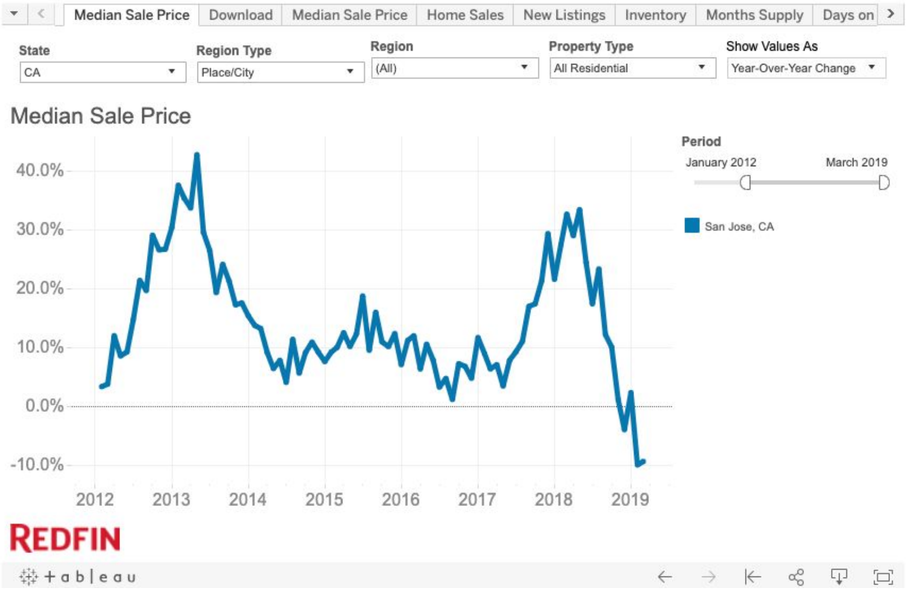
Fig 6: Merged dataframe info

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.

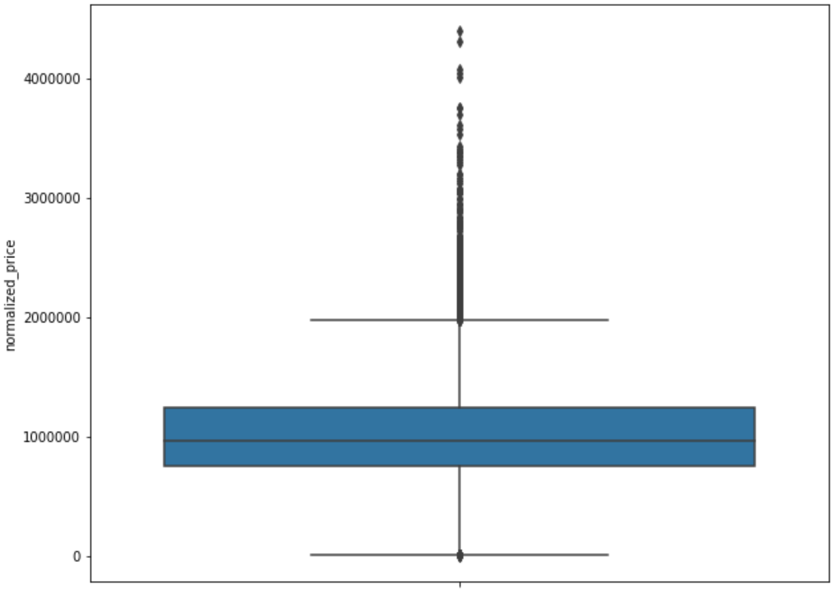


Fig 8: Box plot (adjusted price)

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

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