House Price Prediction in Natural Hazard Prone Areas

Uma Maheswari Raju

Who is the audience?

- Banks and Financial Investors
- Real estate company and marketplace

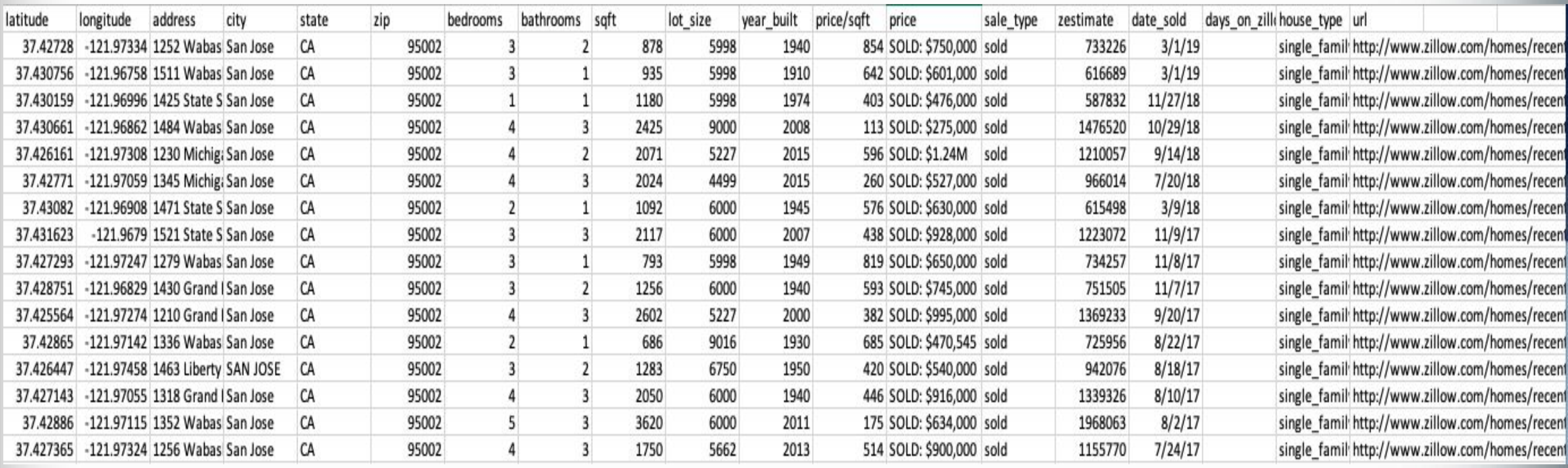
American International Assurance Company, Limited
AIA
AFFINBANK
ALLIANCE BANK
AmBank Group
Bumiputra-Commerce
BANKRAKYAT
Bank Islam
Bank Muamalat
Bank Simpanan Nasional
citibank
EON BANK
Hong Leong
HSBC
ING INSURANCE
Maybank
Malaysia Building Society Berhad
OCBC Bank
PUBLIC BANK
RHB BANK
Scotiabank
Standard Chartered
SOUTHERN Bank Berhad
UNITED OVERSEAS BANK - MALAYSIA
direct access
ABN·AMRO Bank
Bank of America
BANK OF CHINA (MALAYSIA) BERHAD
Cagamas Berhad
BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA
Employees PROVIDENT FUND
POST BY HOMELINK
LHDN
LEMBAGA HASIL DALAM NEGERI

Data Collection

- Zillow Properties Data (San Jose)
  - https://www.zillow.com/homes/san-jose_rb/
- Natural Hazard Data
  - Seismic Hazard Data
    - https://spatialservices.conservation.ca.gov/arcgis/rest/services/CGS_Earthquake_Hazard_Zones/SHP_ZoneInfo/MapServer
  - Fire Hazard Data
    - http://www.fire.ca.gov/fire_prevention/fhsz_maps/FHSZ/santa_clara/San_Jose.pdf

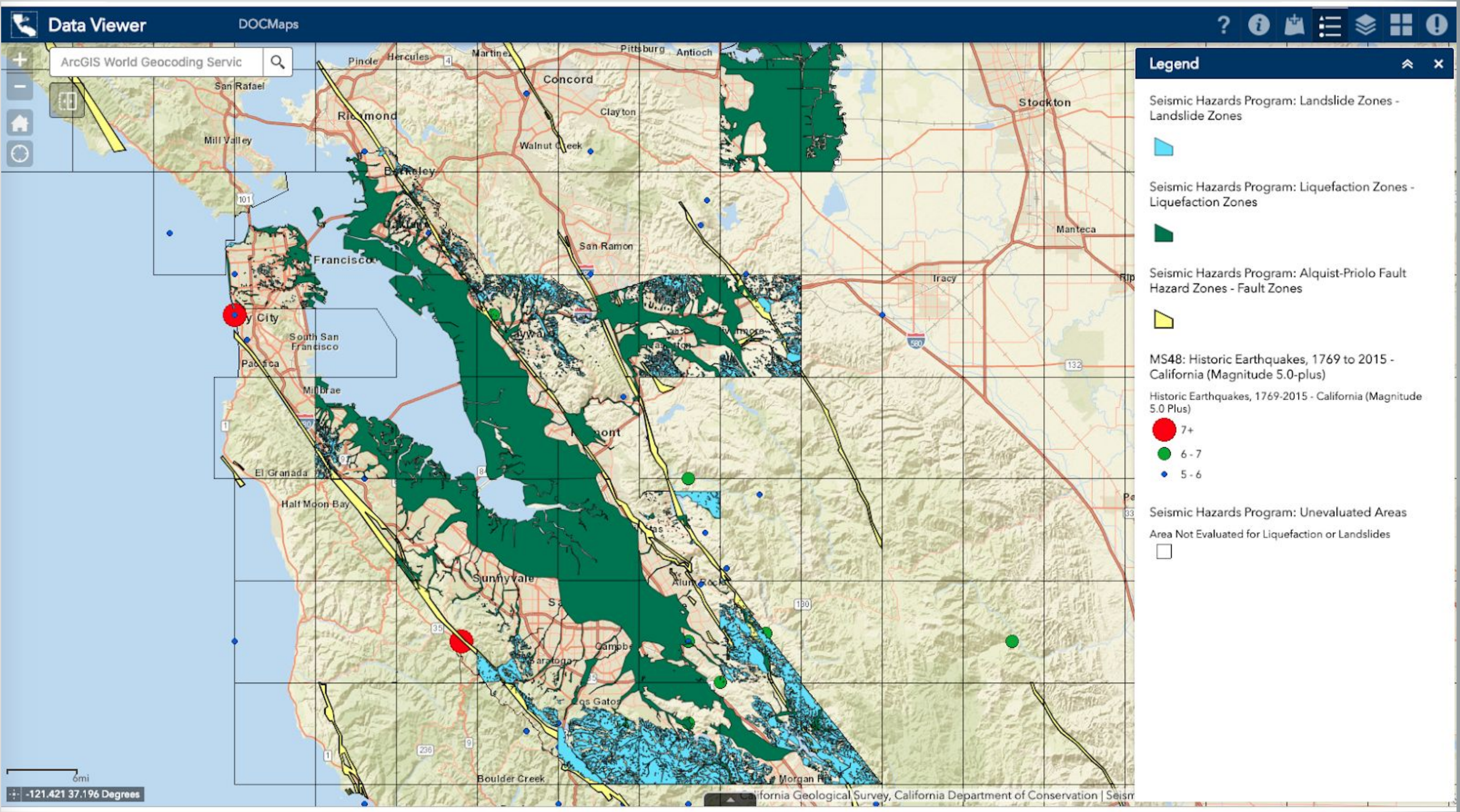
Data Collection – House Data

- Web scraping was done to access Zillow property data using python and web scraping packages such as selenium and BeautifulSoup



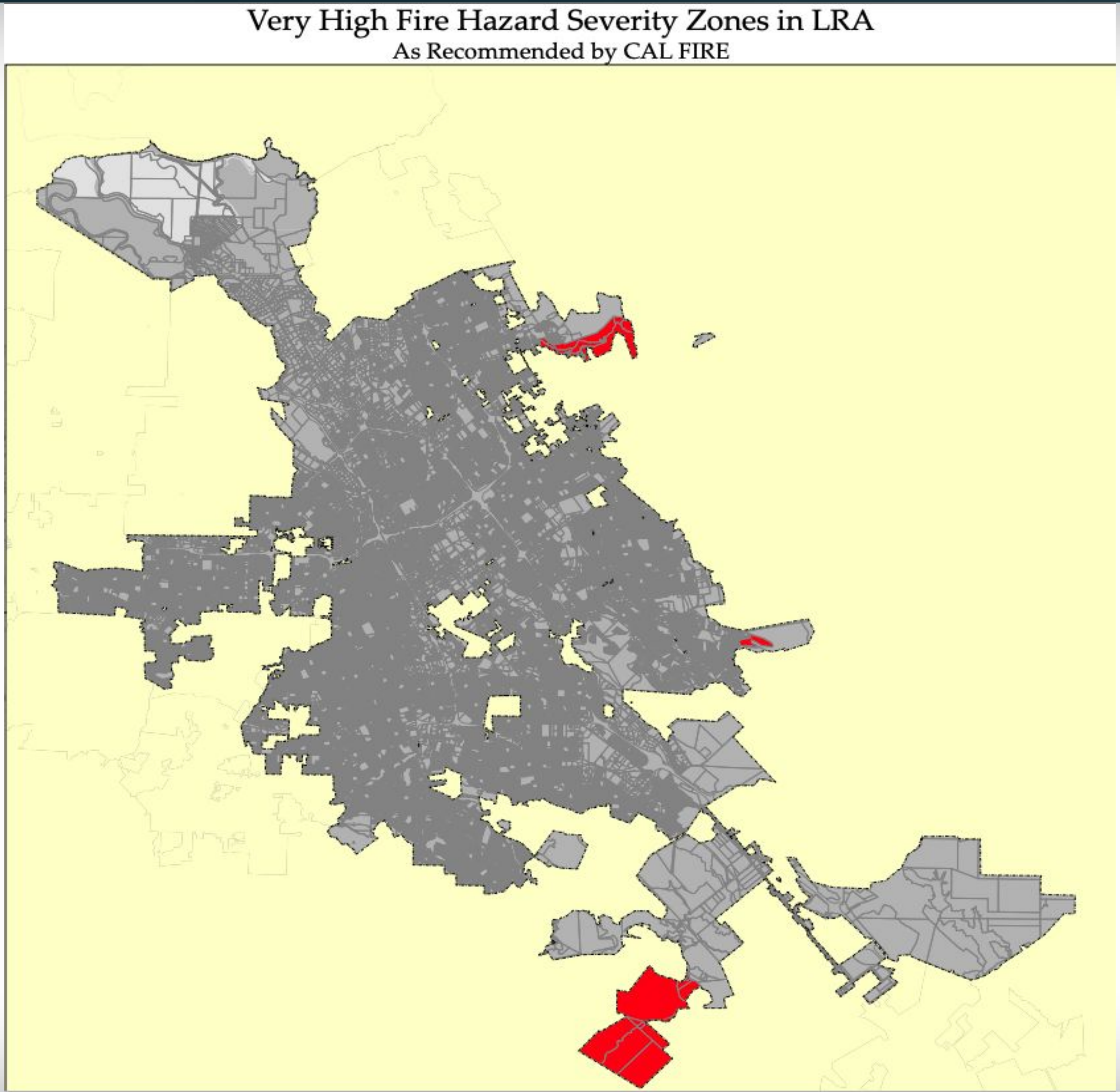
latitude	longitude	address	city	state	zip	bedrooms	bathrooms	sqft	lot_size	year_built	price/sqft	price	sale_type	zestimate	date_sold	days_on_zill	house_type	url
37.42728	-121.97334	1252 Wabas	San Jose	CA	95002	3	2	878	5998	1940	854	SOLD: $750,000	sold	733226	3/1/19		single_famil	http://www.zillow.com/homes/recent
37.430756	-121.96758	1511 Wabas	San Jose	CA	95002	3	1	935	5998	1910	642	SOLD: $601,000	sold	616689	3/1/19		single_famil	http://www.zillow.com/homes/recent
37.430159	-121.96996	1425 State S	San Jose	CA	95002	1	1	1180	5998	1974	403	SOLD: $476,000	sold	587832	11/27/18		single_famil	http://www.zillow.com/homes/recent
37.430661	-121.96862	1484 Wabas	San Jose	CA	95002	4	3	2425	9000	2008	113	SOLD: $275,000	sold	1476520	10/29/18		single_famil	http://www.zillow.com/homes/recent
37.426161	-121.97308	1230 Michig	San Jose	CA	95002	4	2	2071	5227	2015	596	SOLD: $1.24M	sold	1210057	9/14/18		single_famil	http://www.zillow.com/homes/recent
37.42771	-121.97059	1345 Michig	San Jose	CA	95002	4	3	2024	4499	2015	260	SOLD: $527,000	sold	966014	7/20/18		single_famil	http://www.zillow.com/homes/recent
37.43082	-121.96908	1471 State S	San Jose	CA	95002	2	1	1092	6000	1945	576	SOLD: $630,000	sold	615498	3/9/18		single_famil	http://www.zillow.com/homes/recent
37.431623	-121.9679	1521 State S	San Jose	CA	95002	3	3	2117	6000	2007	438	SOLD: $928,000	sold	1223072	11/9/17		single_famil	http://www.zillow.com/homes/recent
37.427293	-121.97247	1279 Wabas	San Jose	CA	95002	3	1	793	5998	1949	819	SOLD: $650,000	sold	734257	11/8/17		single_famil	http://www.zillow.com/homes/recent
37.428751	-121.96829	1430 Grand I	San Jose	CA	95002	3	2	1256	6000	1940	593	SOLD: $745,000	sold	751505	11/7/17		single_famil	http://www.zillow.com/homes/recent
37.425564	-121.97274	1210 Grand I	San Jose	CA	95002	4	3	2602	5227	2000	382	SOLD: $995,000	sold	1369233	9/20/17		single_famil	http://www.zillow.com/homes/recent
37.42865	-121.97142	1336 Wabas	San Jose	CA	95002	2	1	686	9016	1930	685	SOLD: $470,545	sold	725956	8/22/17		single_famil	http://www.zillow.com/homes/recent
37.426447	-121.97458	1463 Liberty	SAN JOSE	CA	95002	3	2	1283	6750	1950	420	SOLD: $540,000	sold	942076	8/18/17		single_famil	http://www.zillow.com/homes/recent
37.427143	-121.97055	1318 Grand I	San Jose	CA	95002	4	3	2050	6000	1940	446	SOLD: $916,000	sold	1339326	8/10/17		single_famil	http://www.zillow.com/homes/recent
37.42886	-121.97115	1352 Wabas	San Jose	CA	95002	5	3	3620	6000	2011	175	SOLD: $634,000	sold	1968063	8/2/17		single_famil	http://www.zillow.com/homes/recent
37.427365	-121.97324	1256 Wabas	San Jose	CA	95002	4	3	1750	5662	2013	514	SOLD: $900,000	sold	1155770	7/24/17		single_famil	http://www.zillow.com/homes/recent

Data Collection – Seismic Hazard


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
Walnut Creek
Richmond
Berkeley
San Ramon
Francisco
Stockton
Manteca
Tracy
South San Francisco
Pacifica
Millbrae
El Granada
Half Moon Bay
Sunnyvale
Saratoga
Campbell
Los Gatos
Boulder Creek
Morgan Hill
6mi
-121.421 37.196 Degrees
California Geological Survey, California Department of Conservation | Seism

Source: California Geological Survey (CGS)

Data Collection – Fire Hazard


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

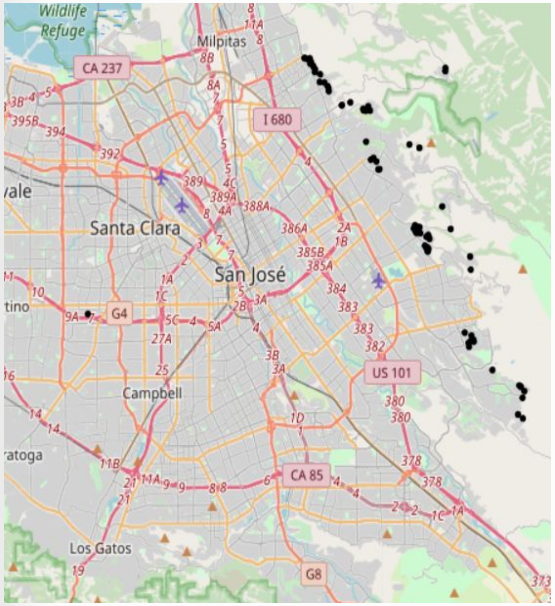
Source: CALFIRE

Data Wrangling

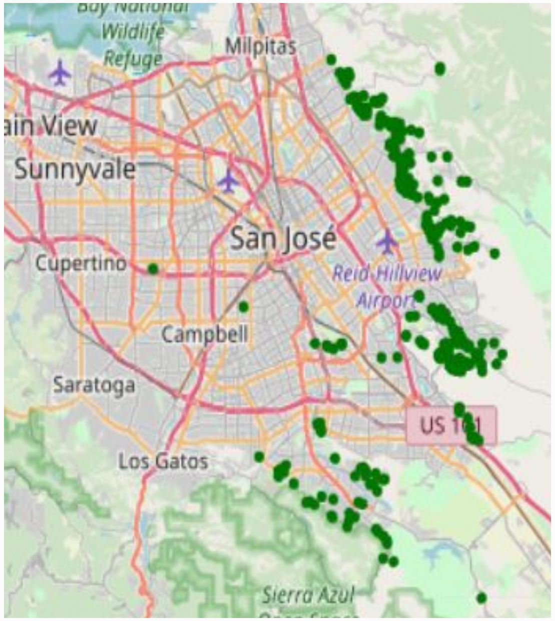
- San Jose City, California
- House sold between 2016-2019
- 13993 observations and 7 features
- All data frames were merged and duplicates were removed
- Different formats were corrected
- Error data were corrected
- Data types were corrected
- Missing data and outliers were handled

Exploratory Data Analysis (EDA)

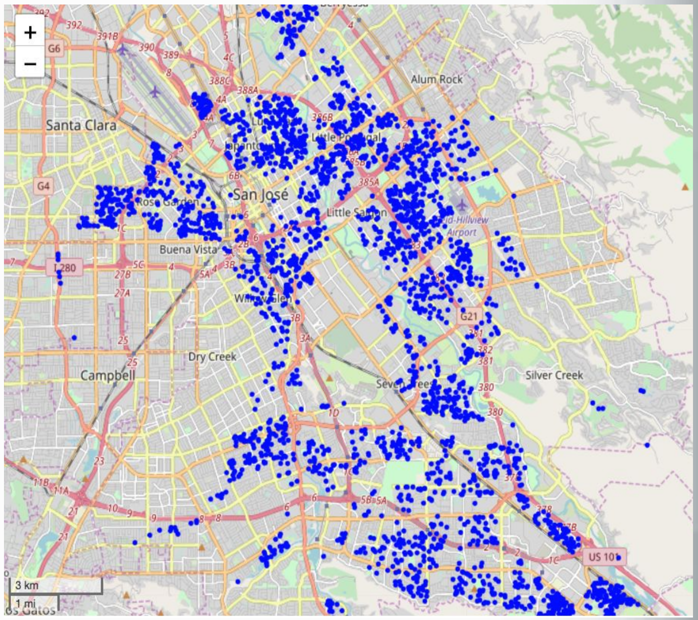
- **Where are hazards prone areas? Which hazard is the most common and least common in city? How are they distributed in the city?**



Fault zone



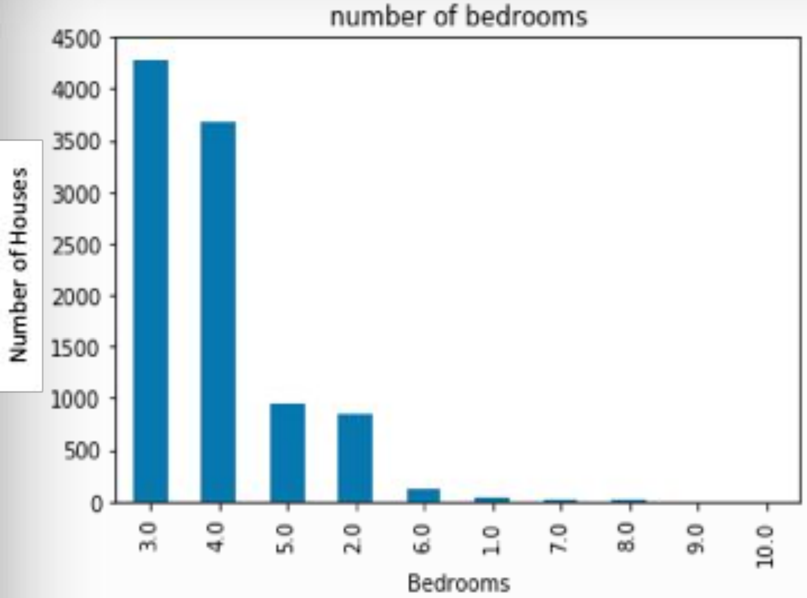
Landslide

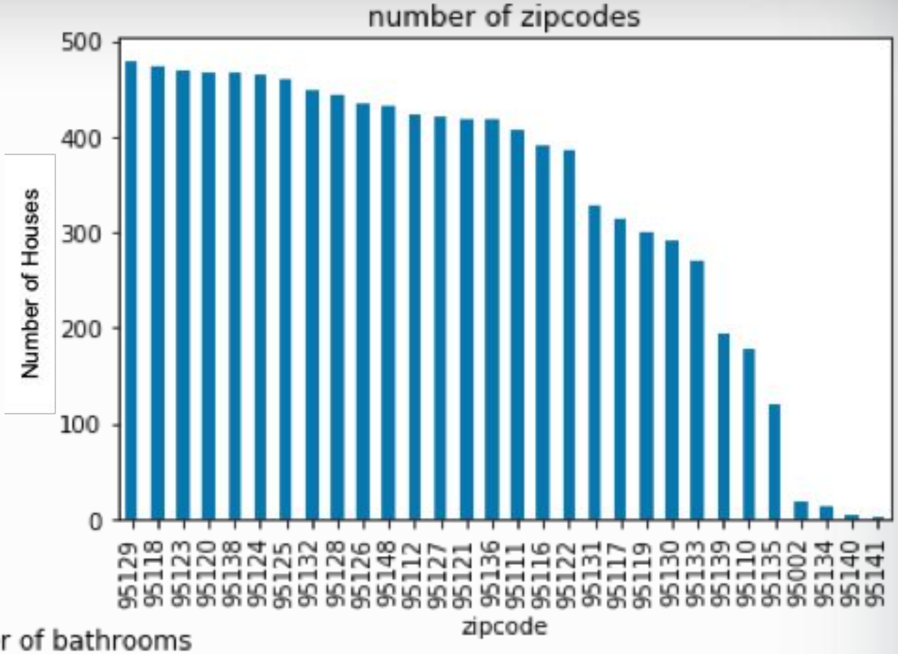


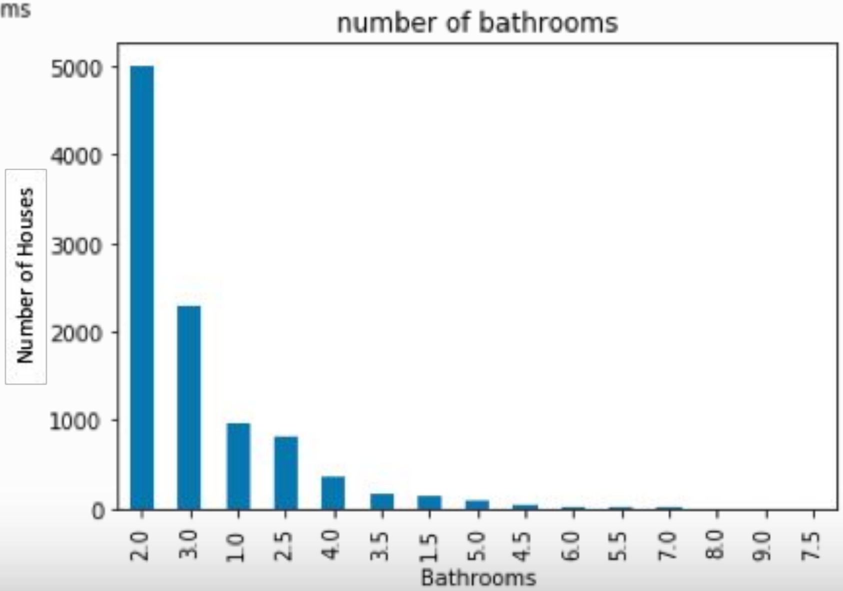
Liquefaction

EDA

- Popular homes among buyers

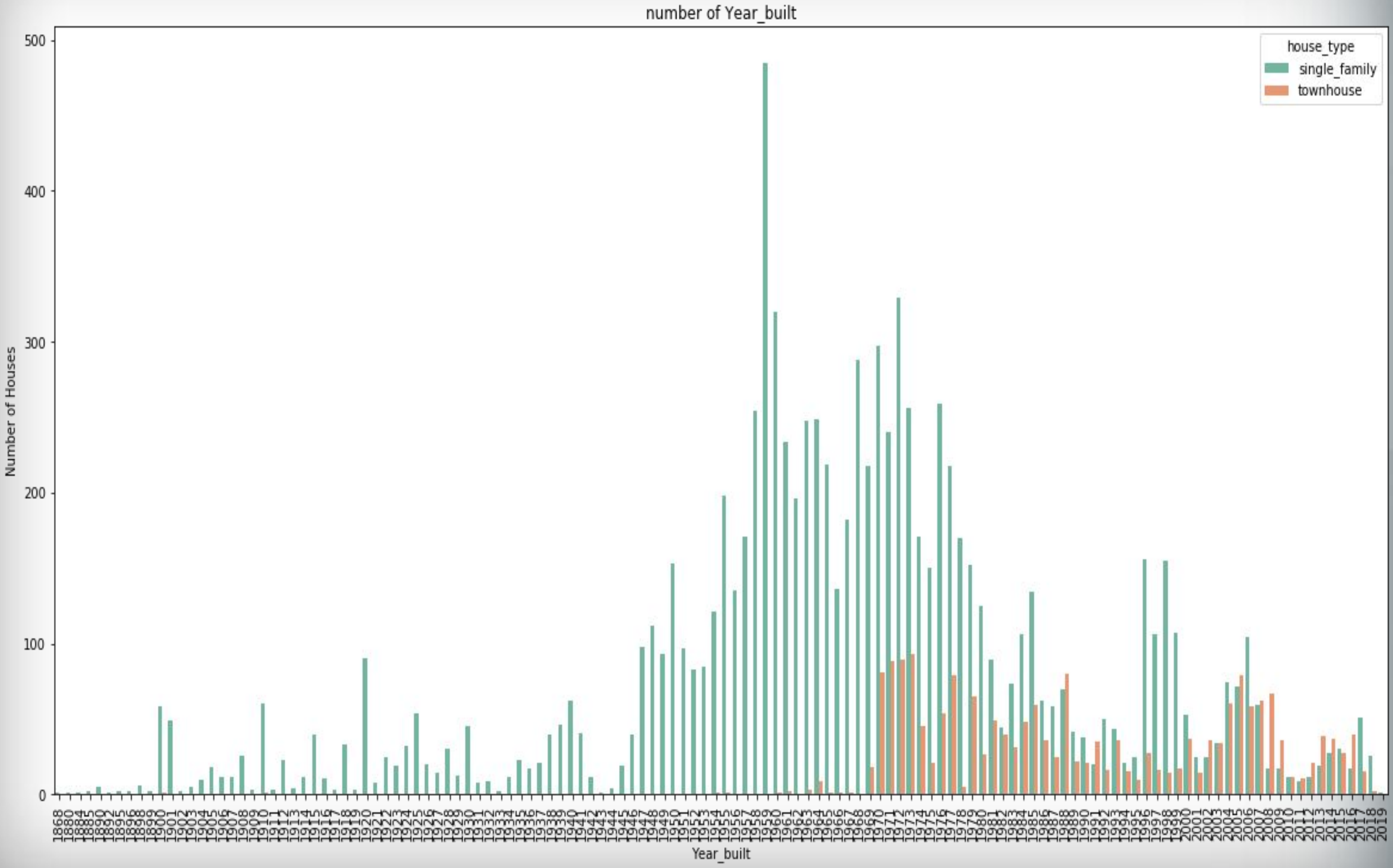






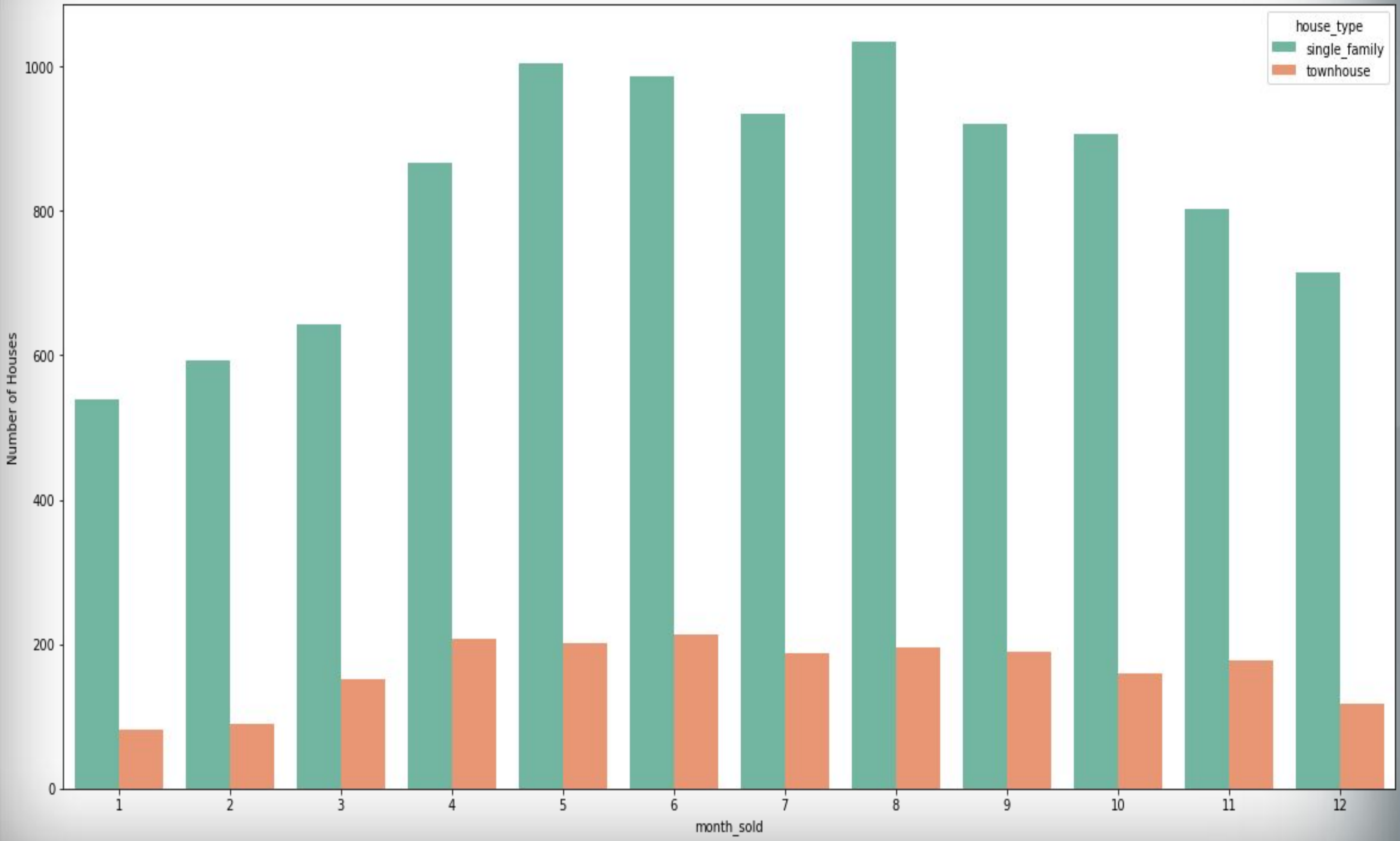
EDA

Which year built is the most common ?



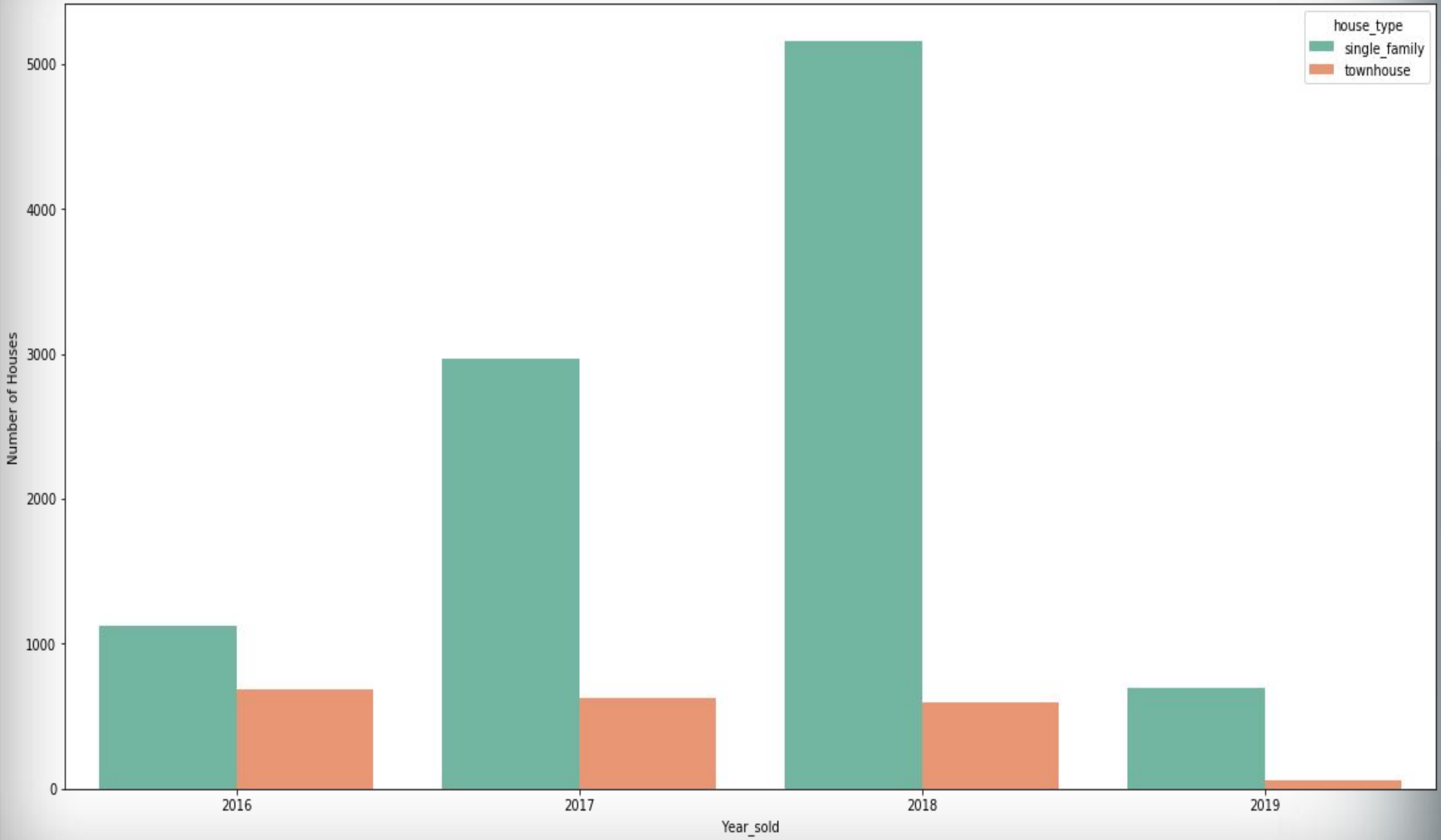
EDA

Which month is popular for selling house?



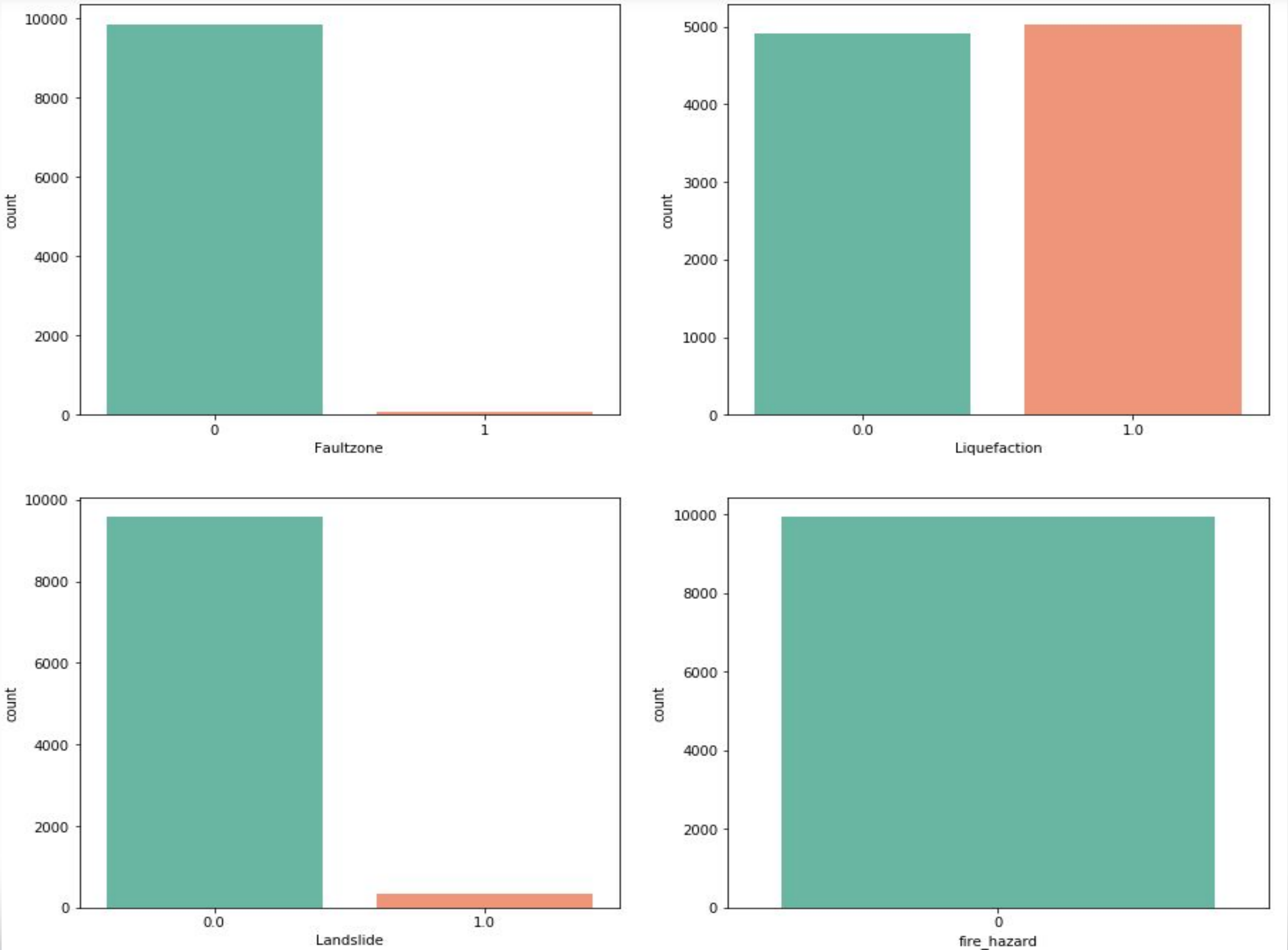
EDA

Which year was most house sold?



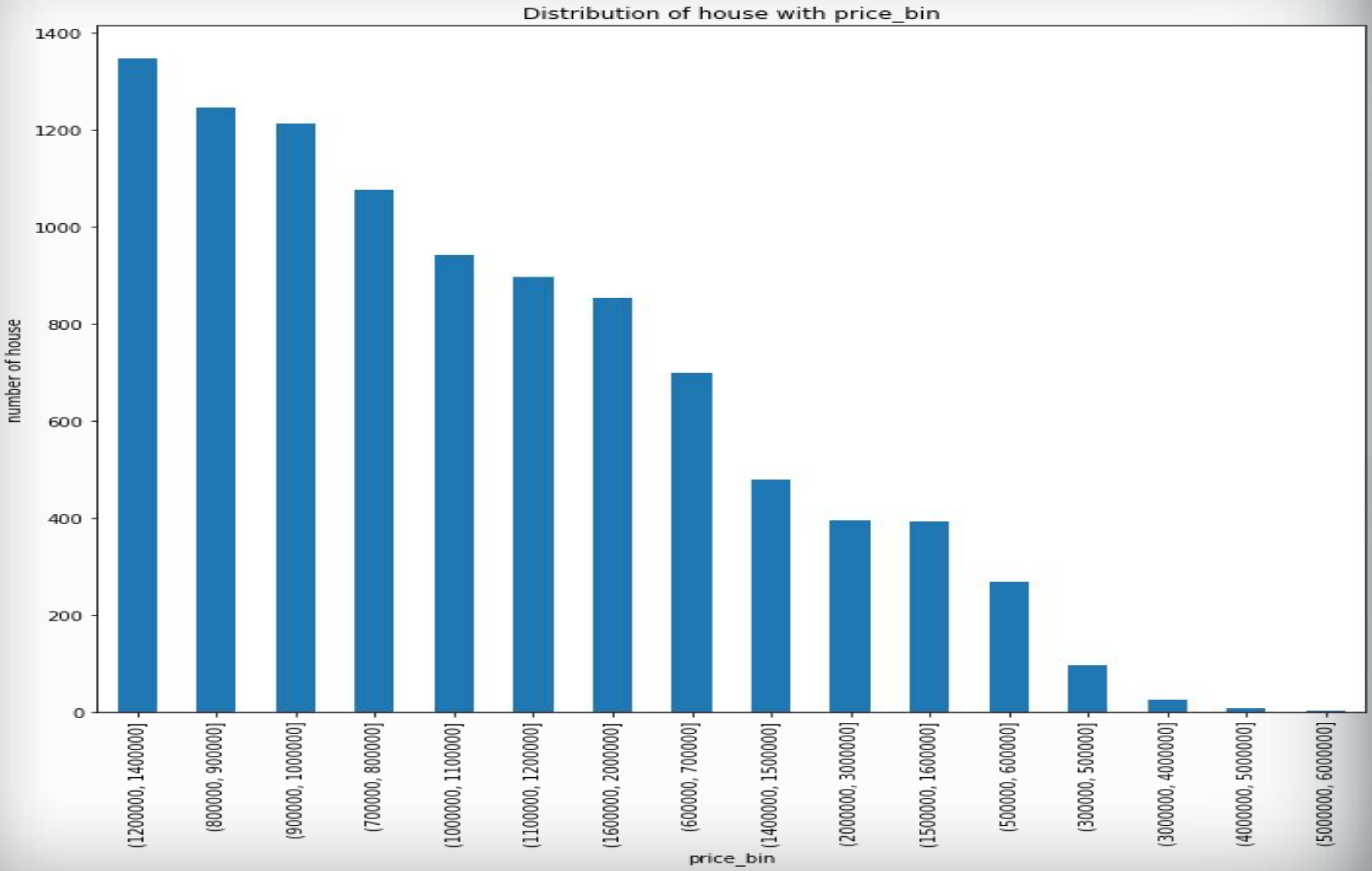
EDA

Which hazard is most common?



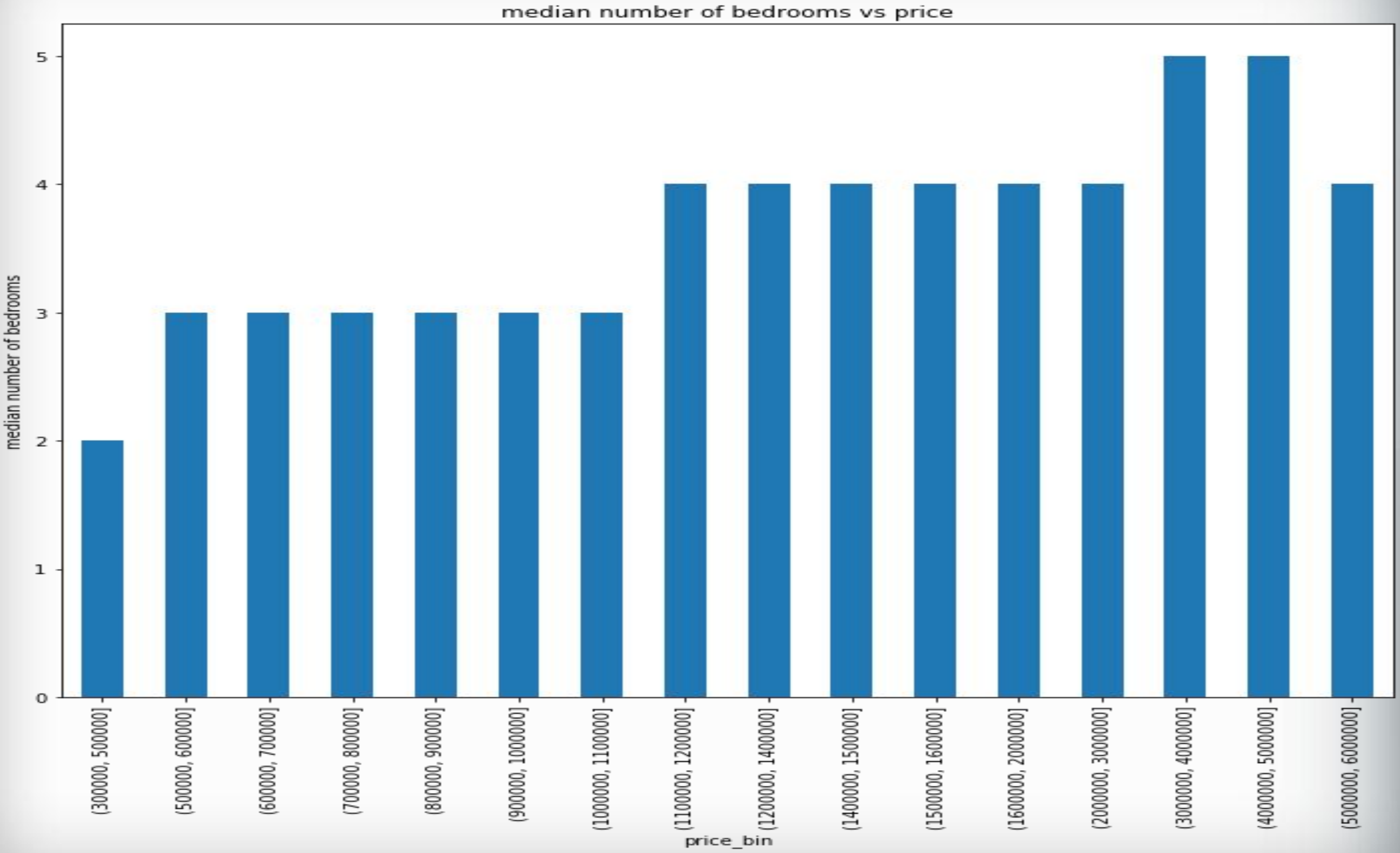
EDA

How many number of houses in different price bin?



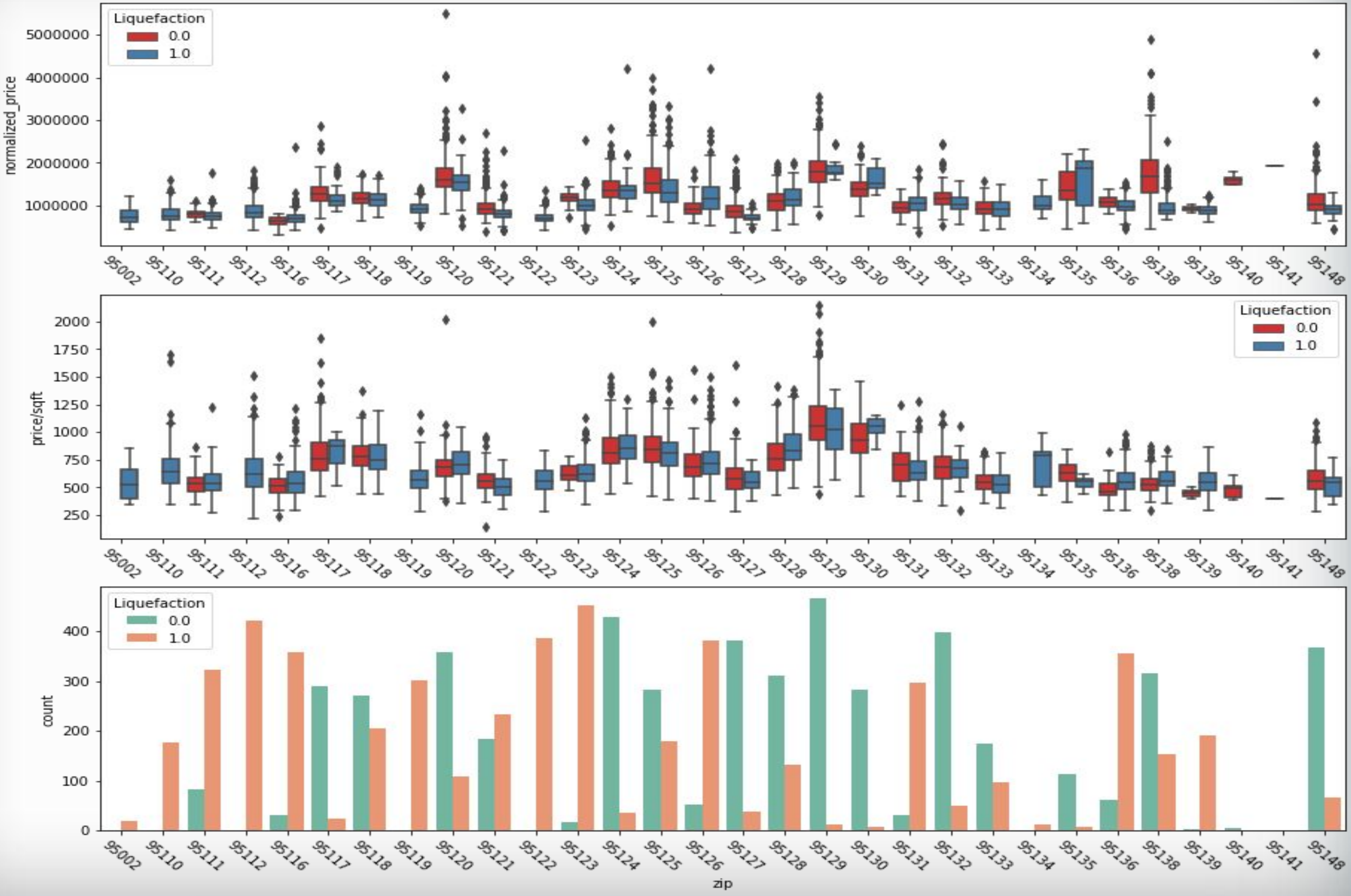
EDA

How many number of bedrooms in different price bin?



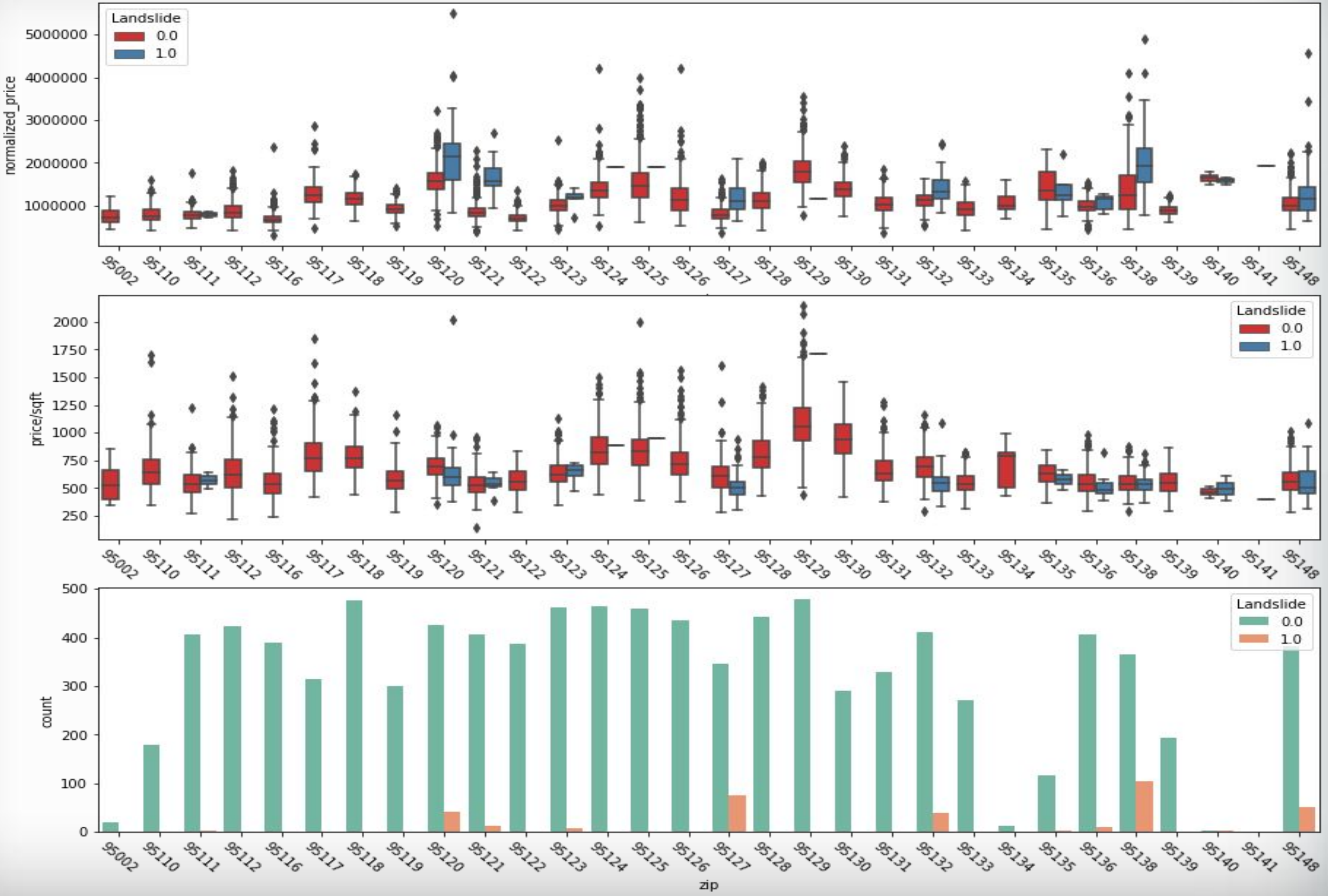
EDA

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



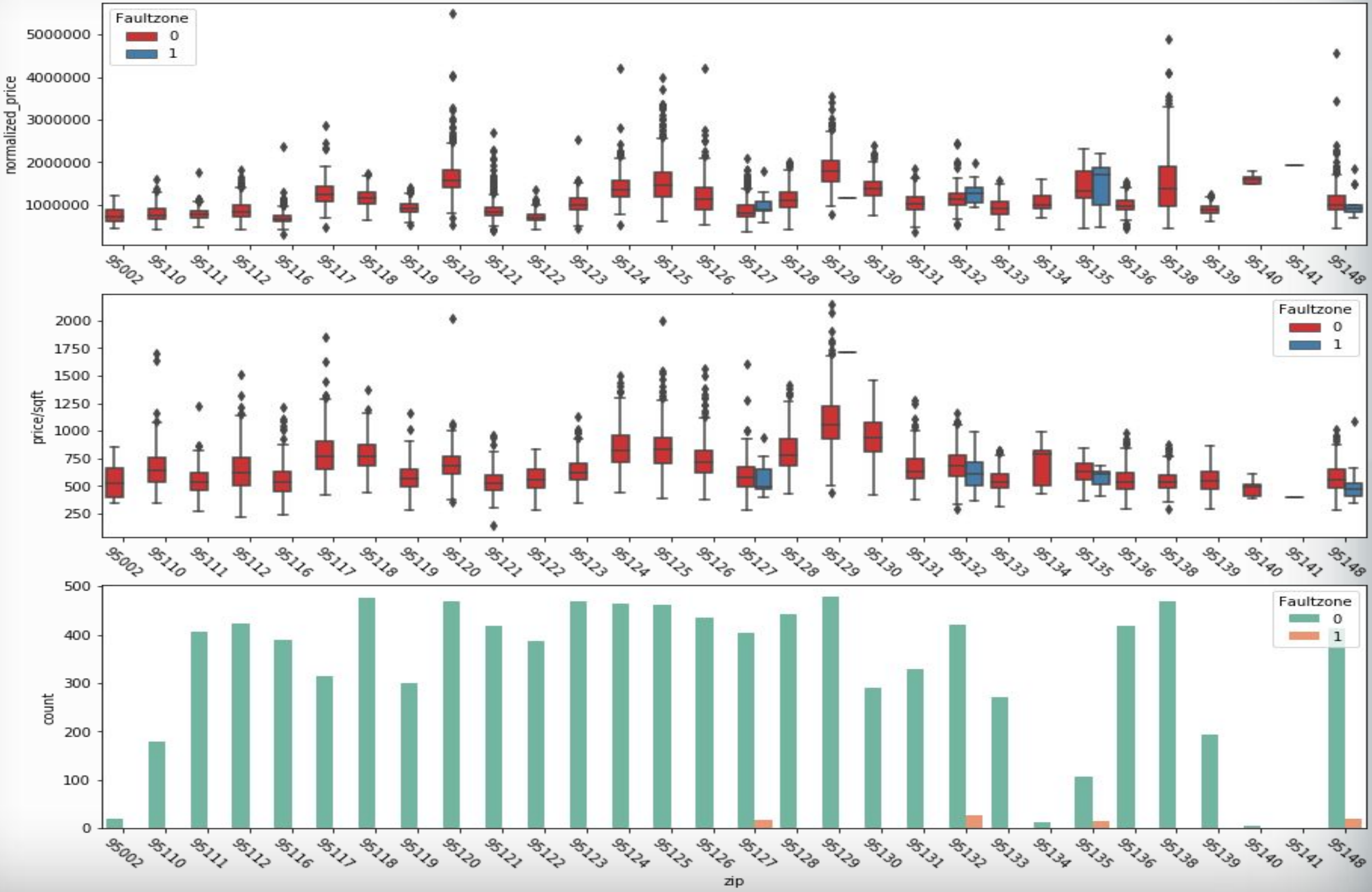
EDA

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



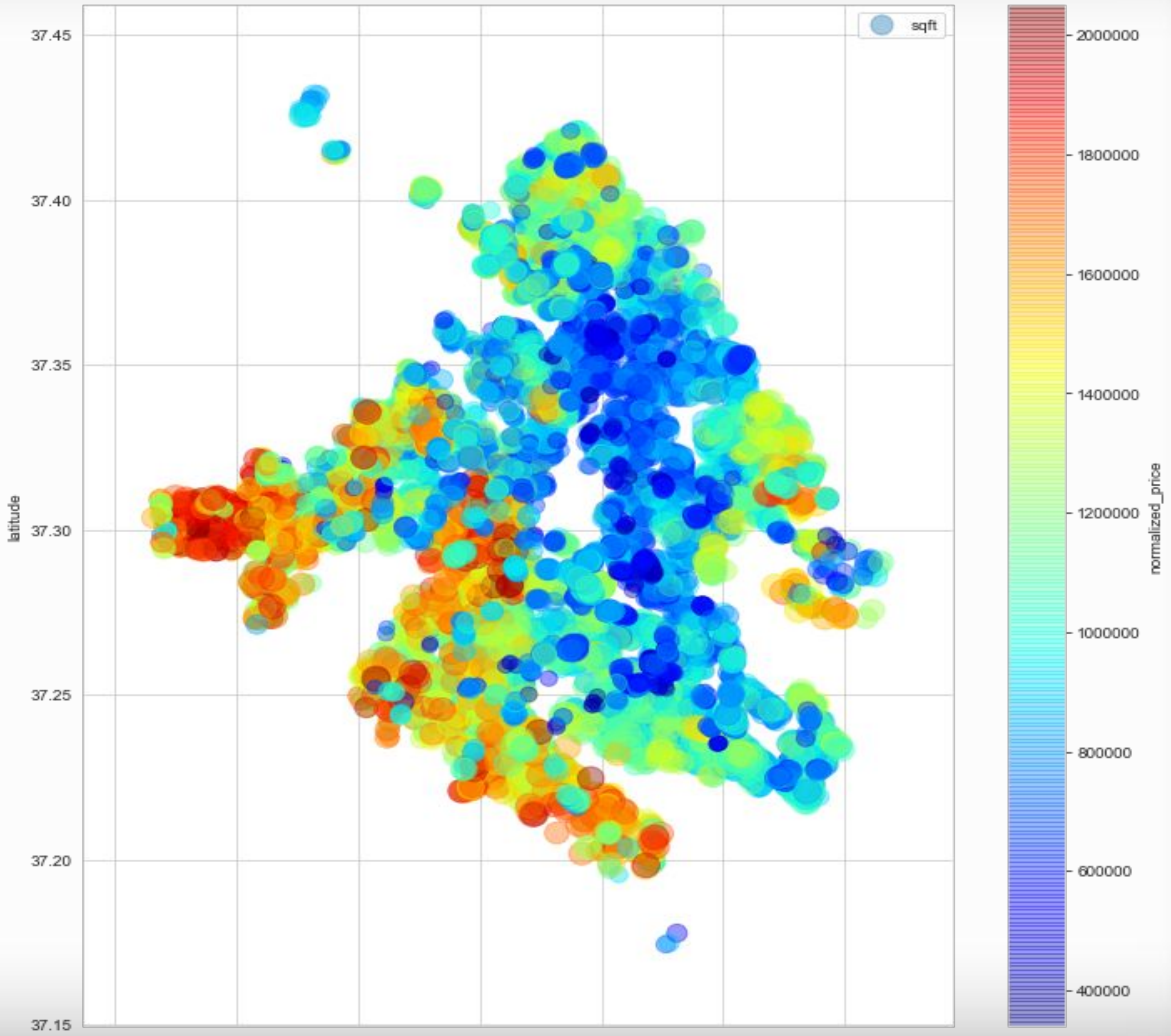
EDA

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

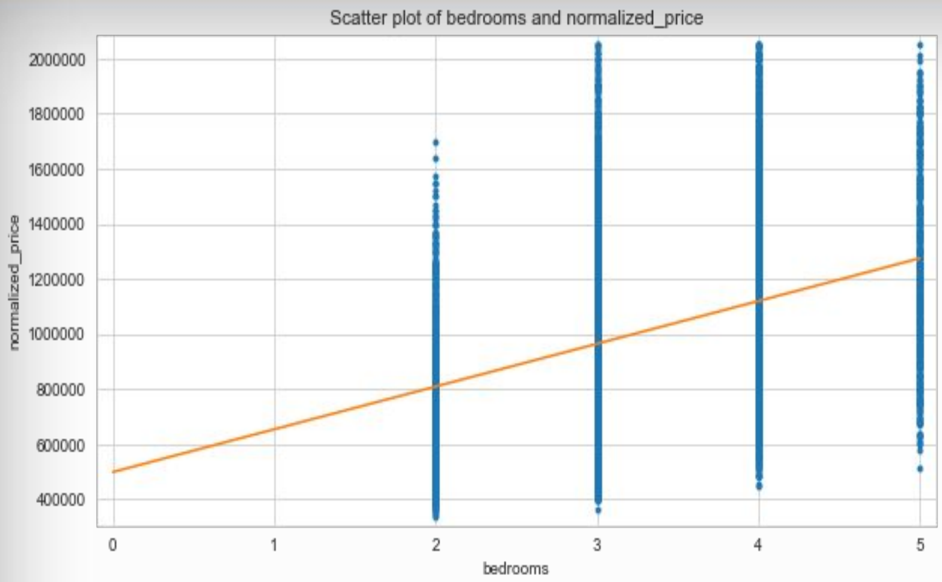


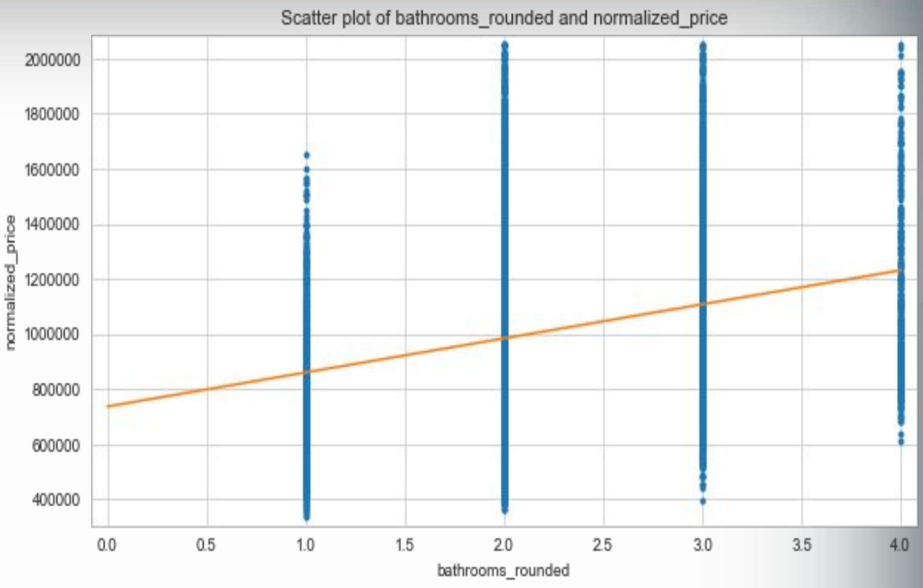
EDA

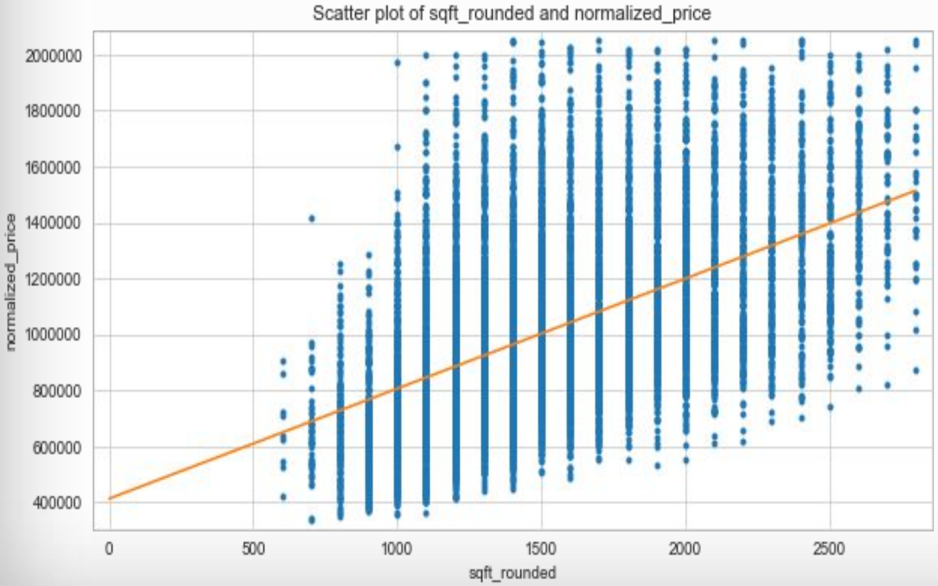
Price distribution in geospatial frame

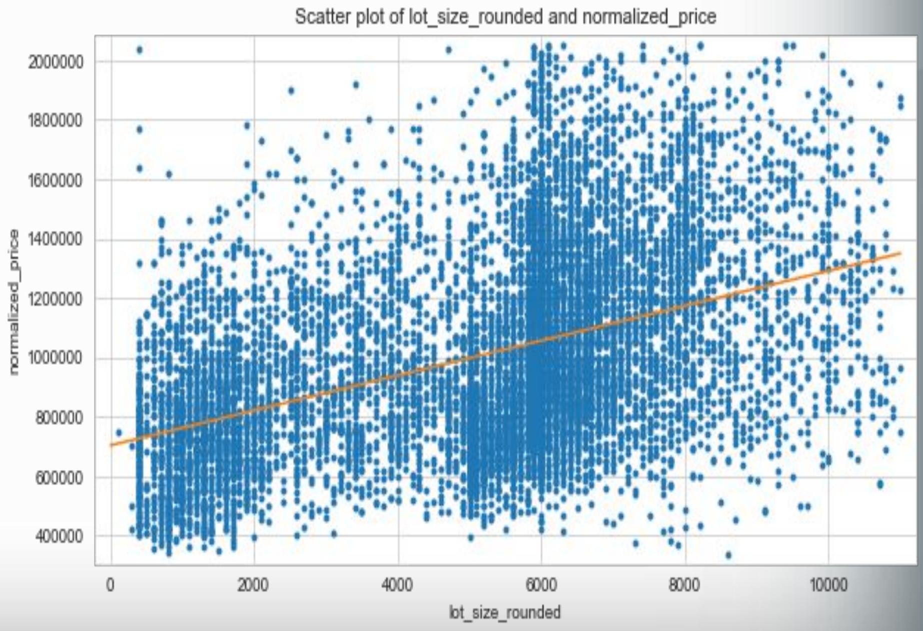


Correlation - Features and Price

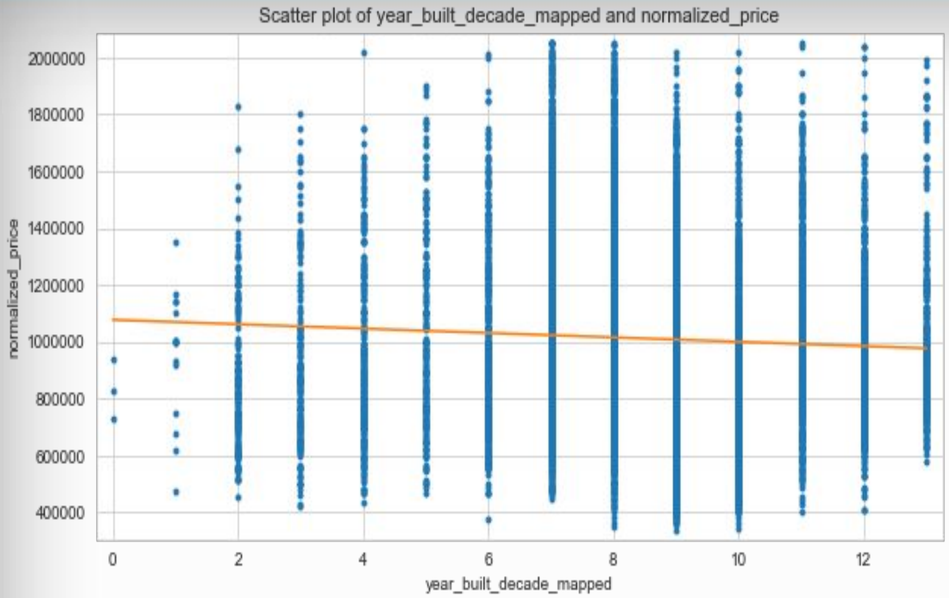


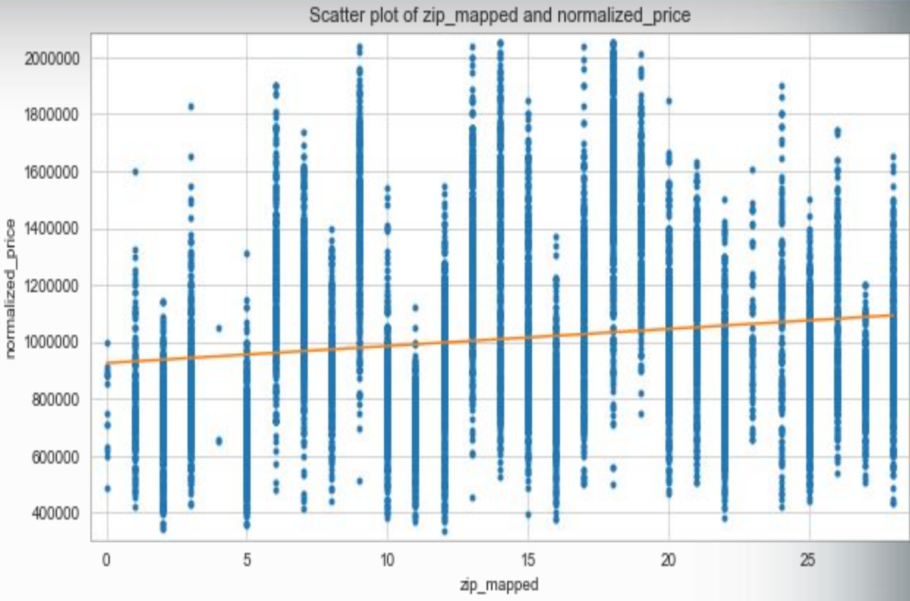


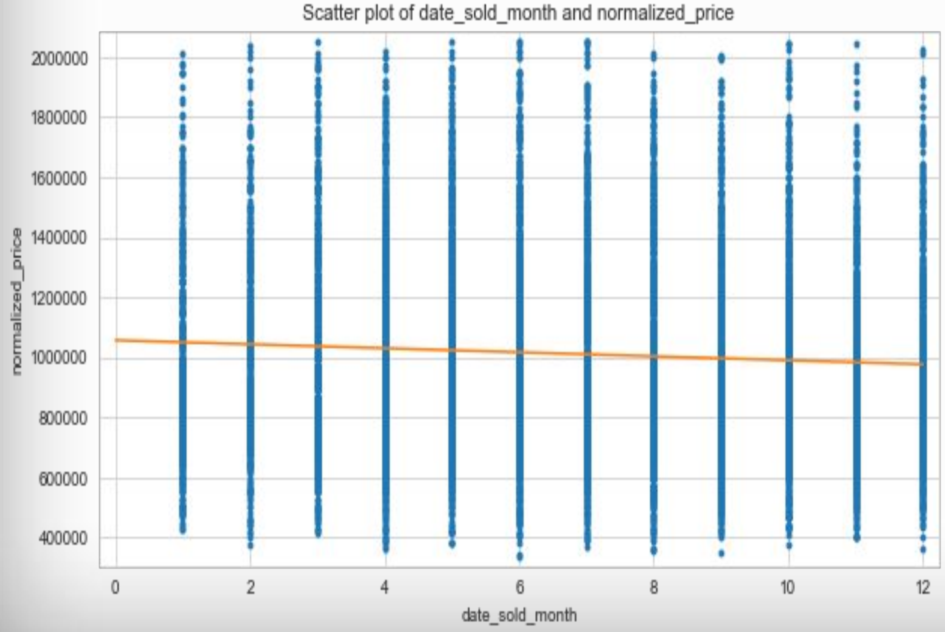


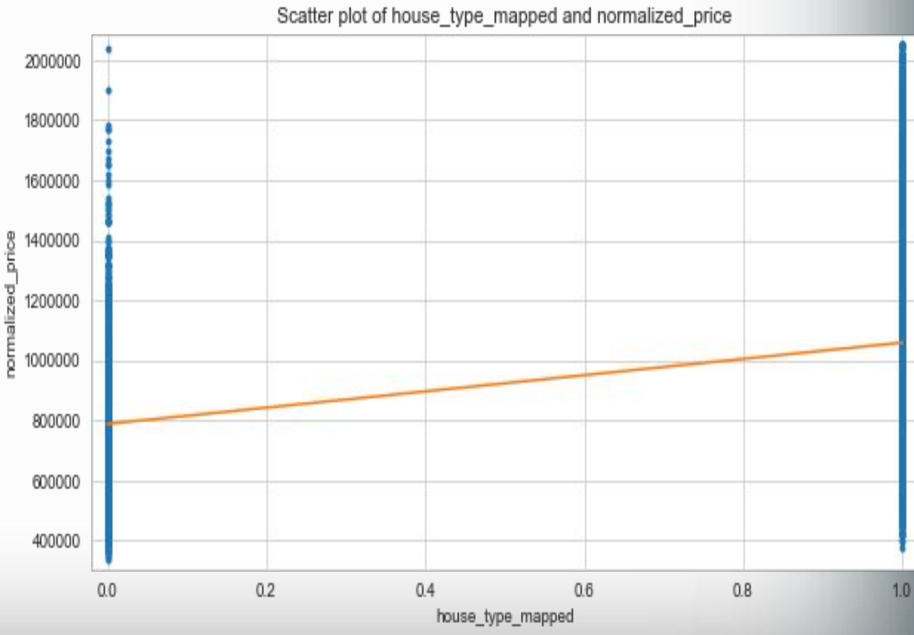


Correlation - Features and Price

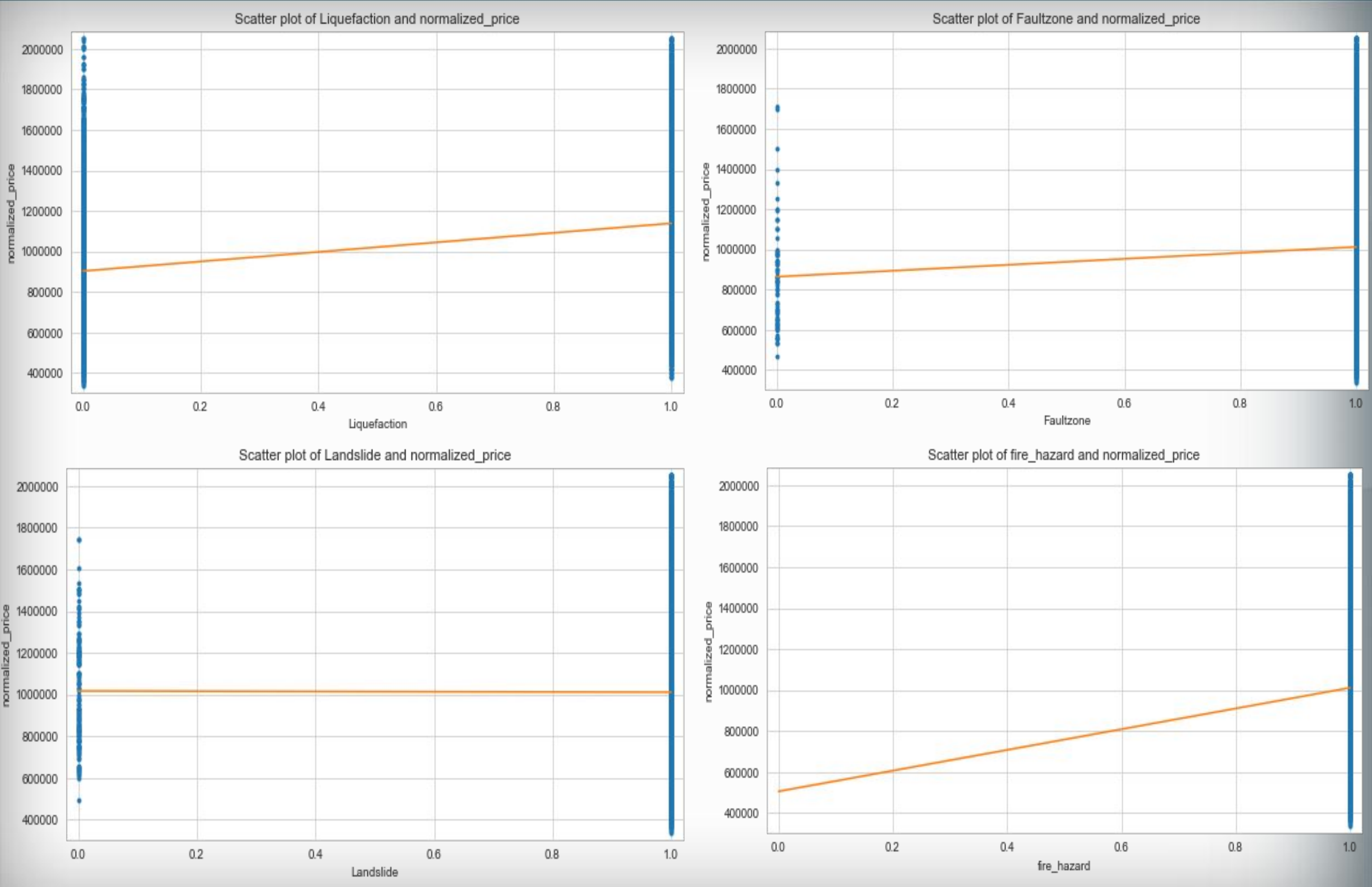

Scatter plot of year_built_decade_mapped and normalized_price


Scatter plot of zip_mapped and normalized_price


Scatter plot of date_sold_month and normalized_price


Scatter plot of house_type_mapped and normalized_price

Correlation - Features and Price



Scatter plot of Liquefaction and normalized_price

Scatter plot of Faultzone and normalized_price

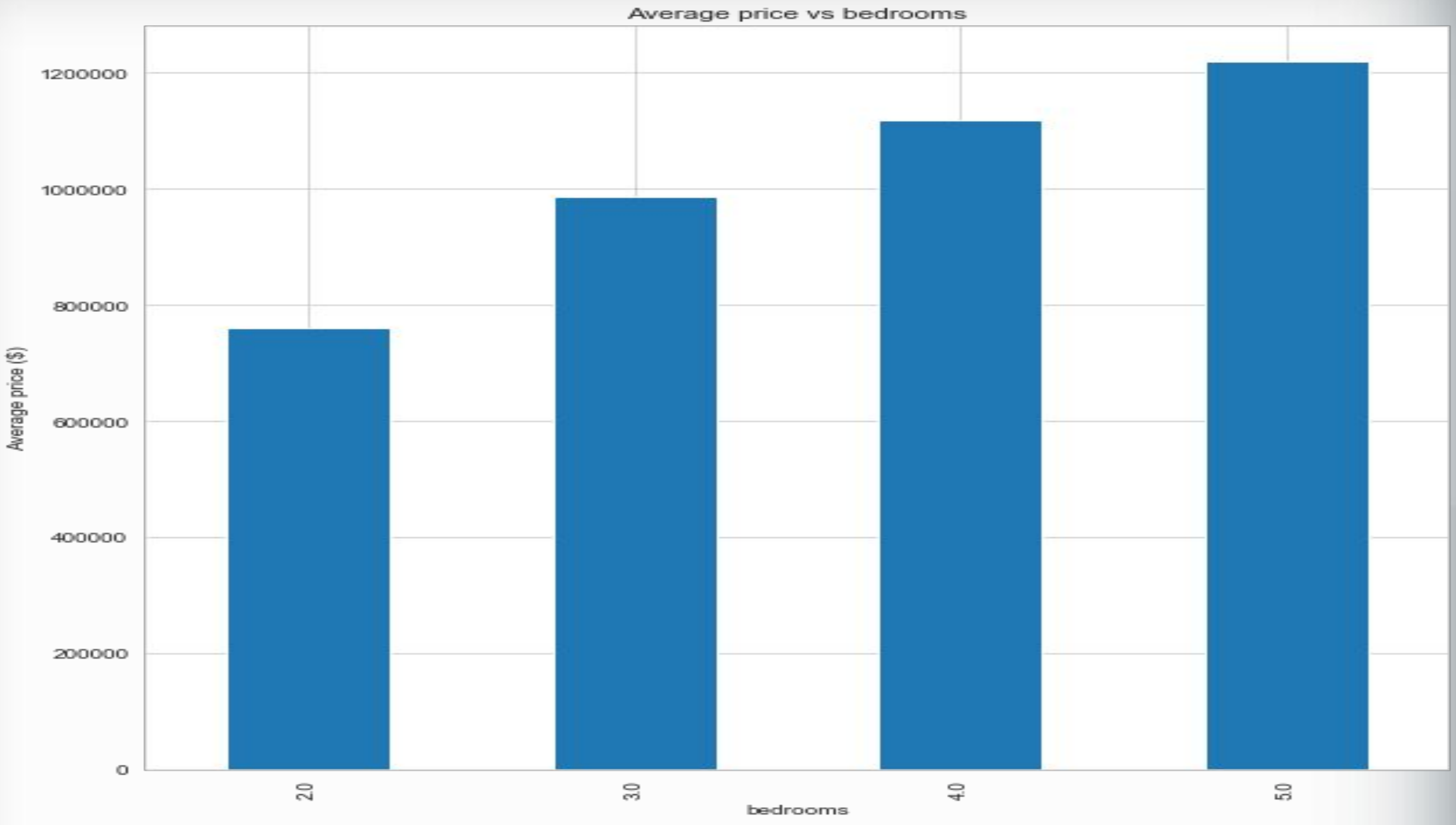
Scatter plot of Landslide and normalized_price

Scatter plot of fire_hazard and normalized_price

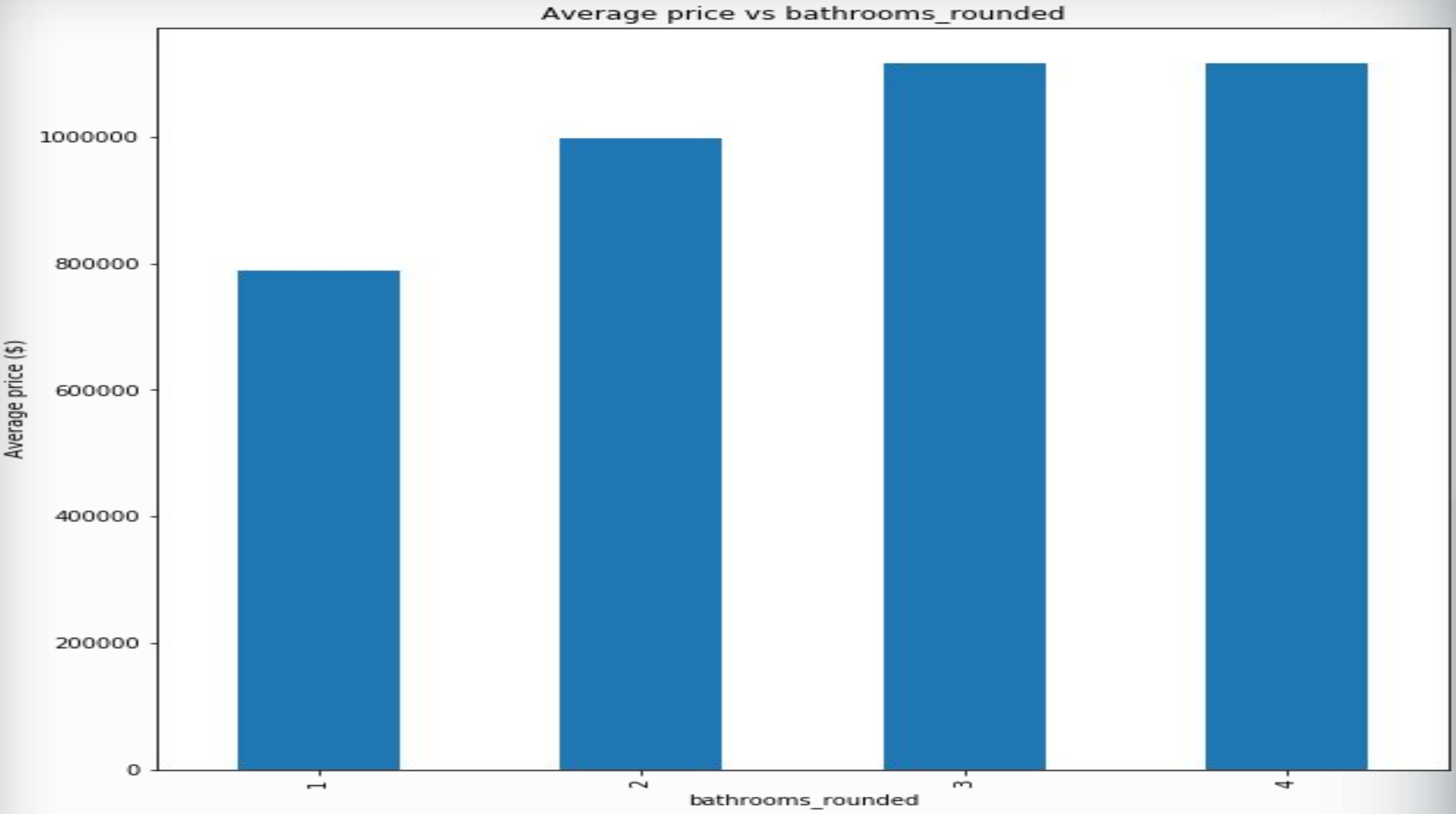
Correlation Coefficients

Sqft	0.521310
Lot size	0.428212
Bedrooms	0.368206
Liquefaction	0.361638
House type	0.317384
Bathrooms	0.250558
Zip code	0.139833
Fault zone	0.036062
Month_sold	-0.065022
Year built	-0.052772
Landslide	-0.002046
Fire hazard	NaN

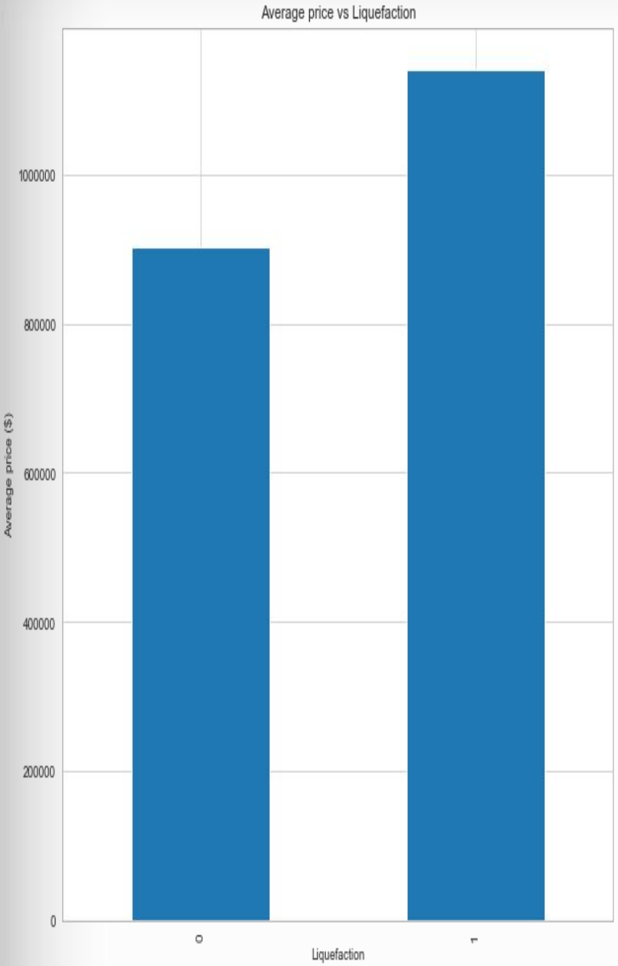
Data Story


Average price vs bedrooms
Average price ($)
1200000
1000000
800000
600000
400000
200000
0
2.0
3.0
4.0
5.0
bedrooms

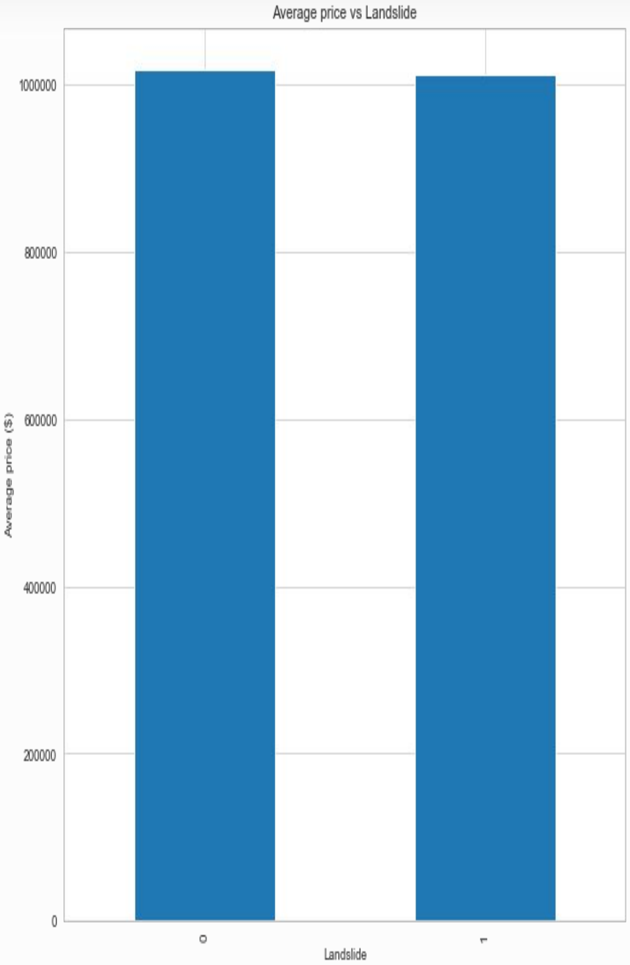
Data Story


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

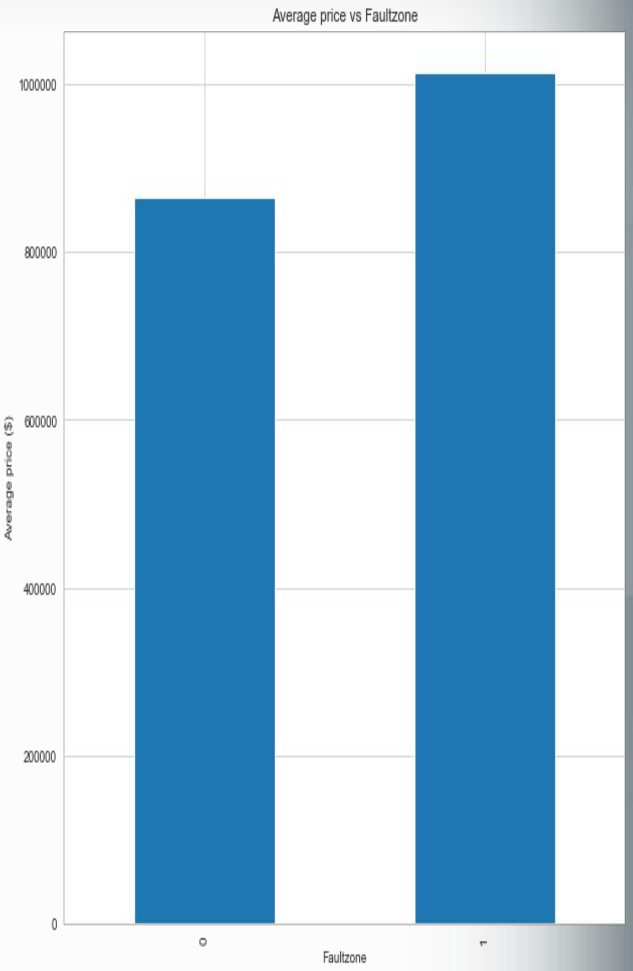
Data Story



Liquefaction

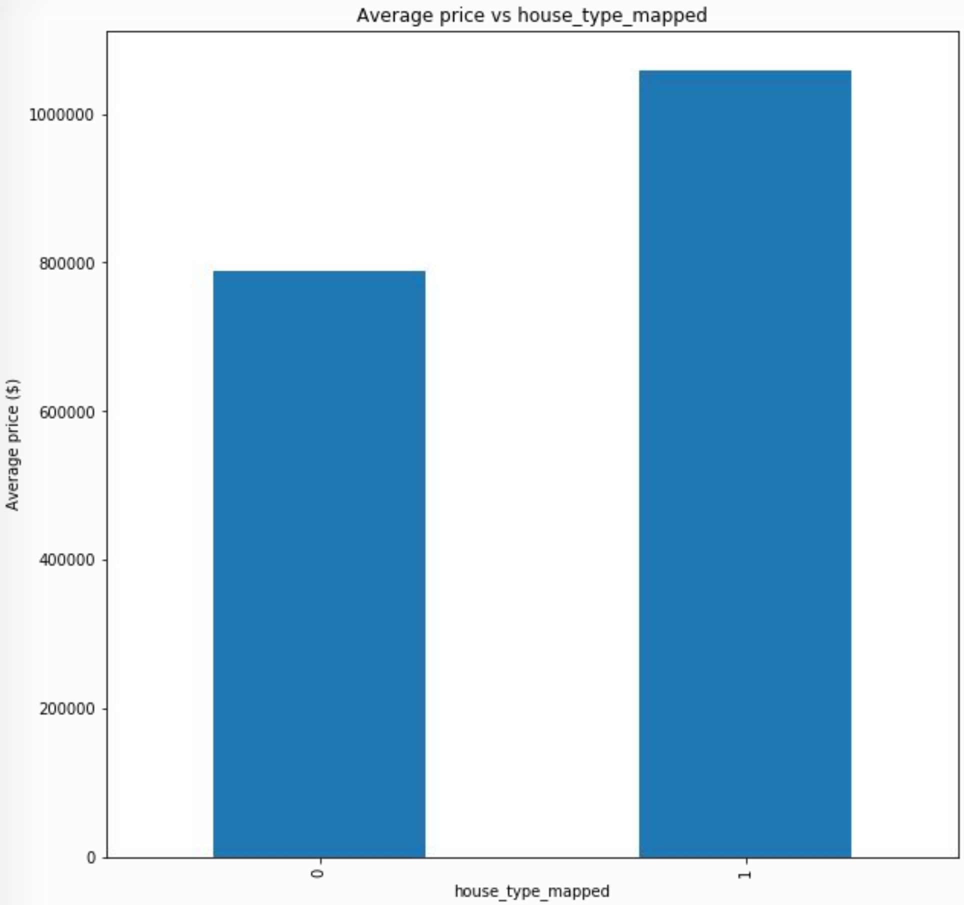


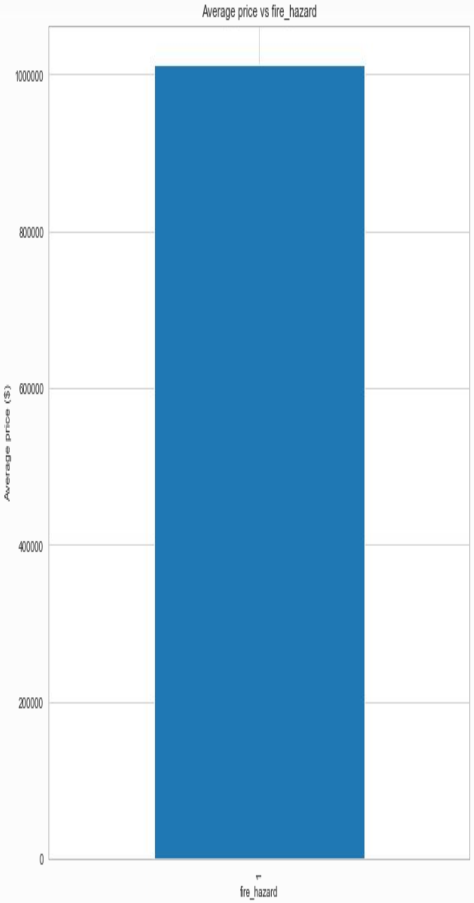
Landslide



Fault zone

Data Story





Fire hazard

Inferential Statistics

- H0: There is no significant correlation between features and price
- Ha: There is a correlation between features and price
- The.
- Features bedrooms, bathrooms, sqft, lot size, liquefaction, fault zone, house type, and zip code, p value is less than level of significance 0.05 which suggested significant correlation between above features and price.
- Other features such as year built, month sold and landslide, p value was greater than 0.05 which suggested no significant correlation between those features and price.

Machine Learning Models

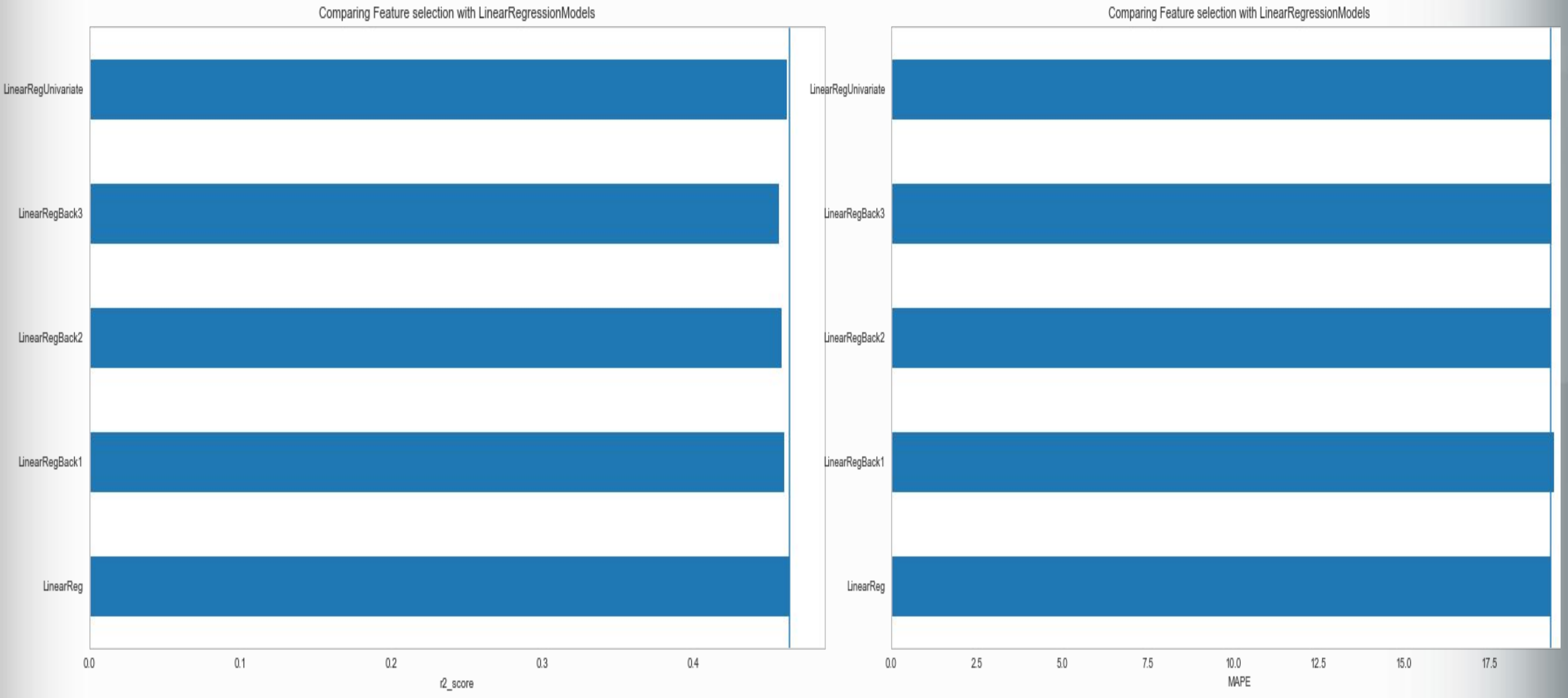
- Linear Regression
- Decision Tree Regressor
- Gradient Boosting Regressor
- Random Forest Regressor

Metrics

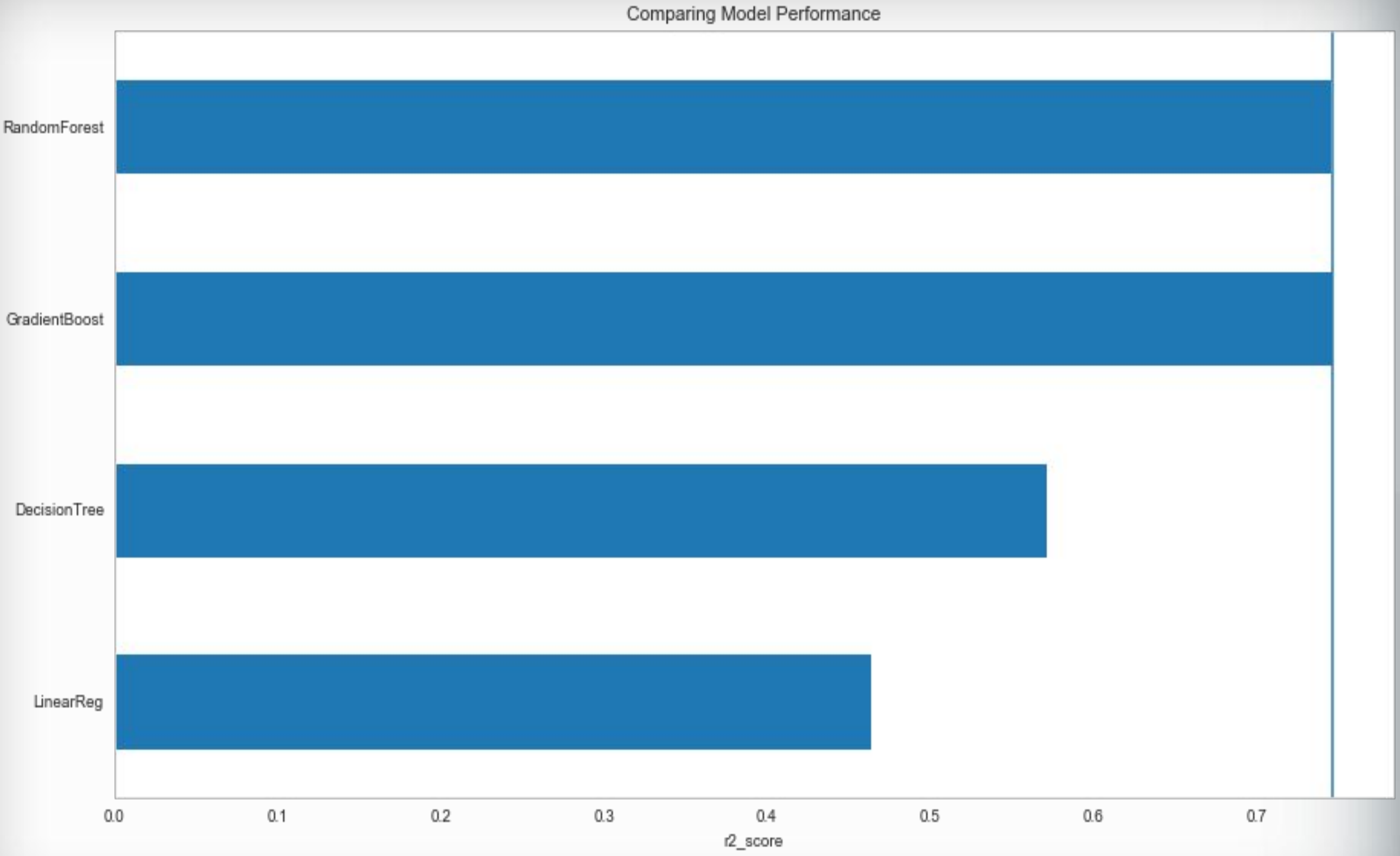
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R2_score
- Mean Absolute Error (MAE)
- Mean Absolute Percent Error (MAPE)

Feature Selection

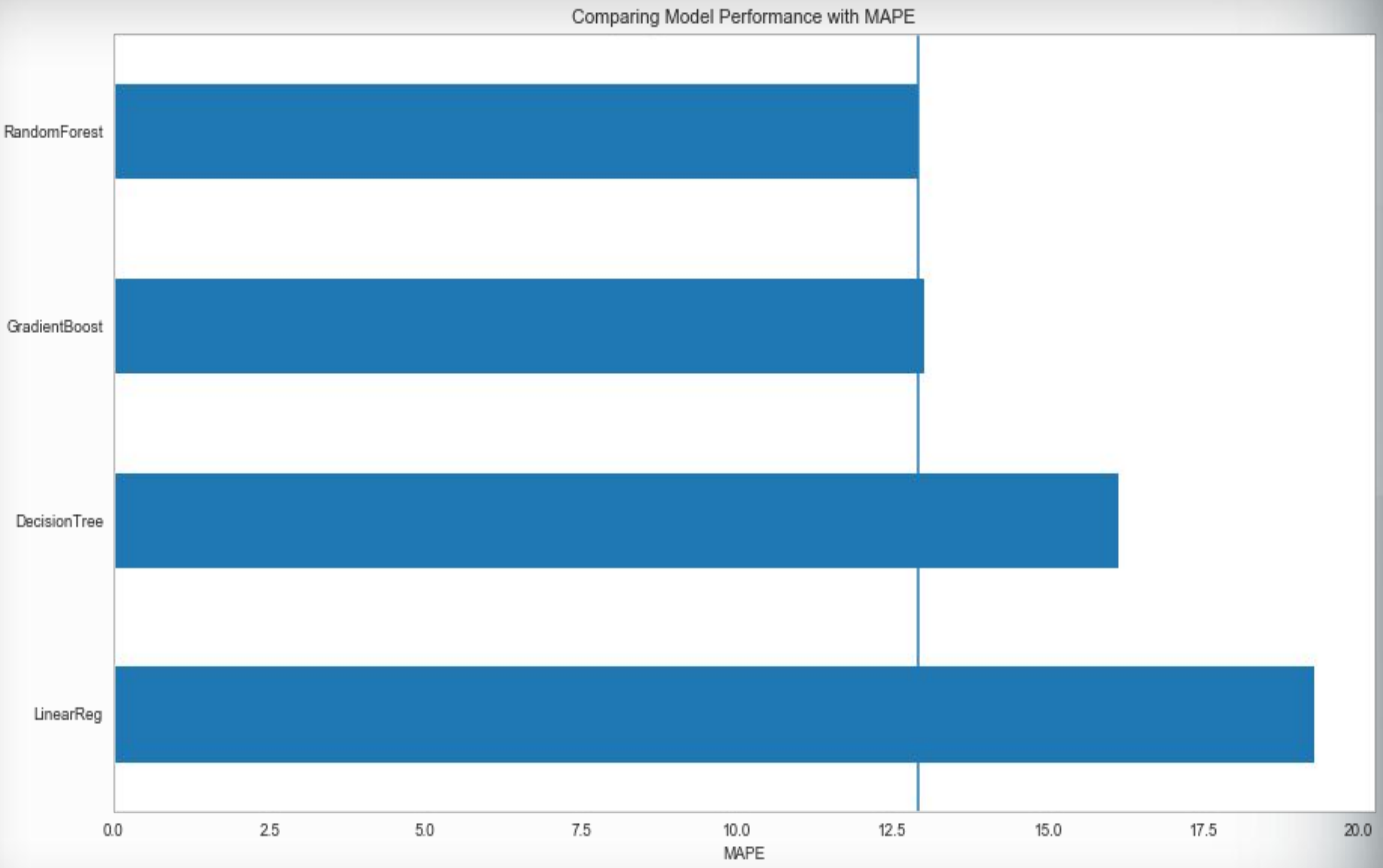
- Backward Elimination
- Univariate Elimination



Comparison of Regressor Models

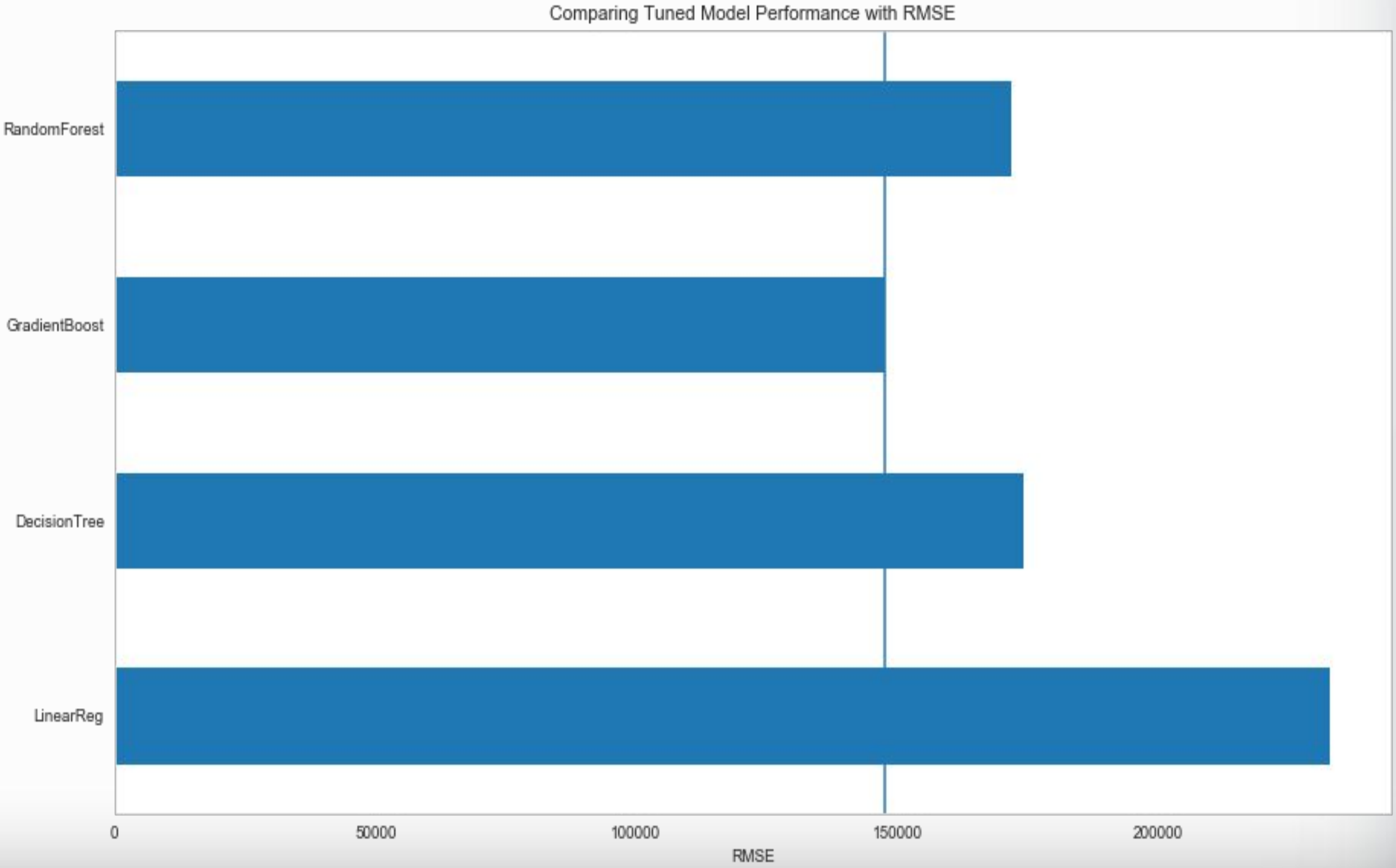


Comparison of Regressor Models

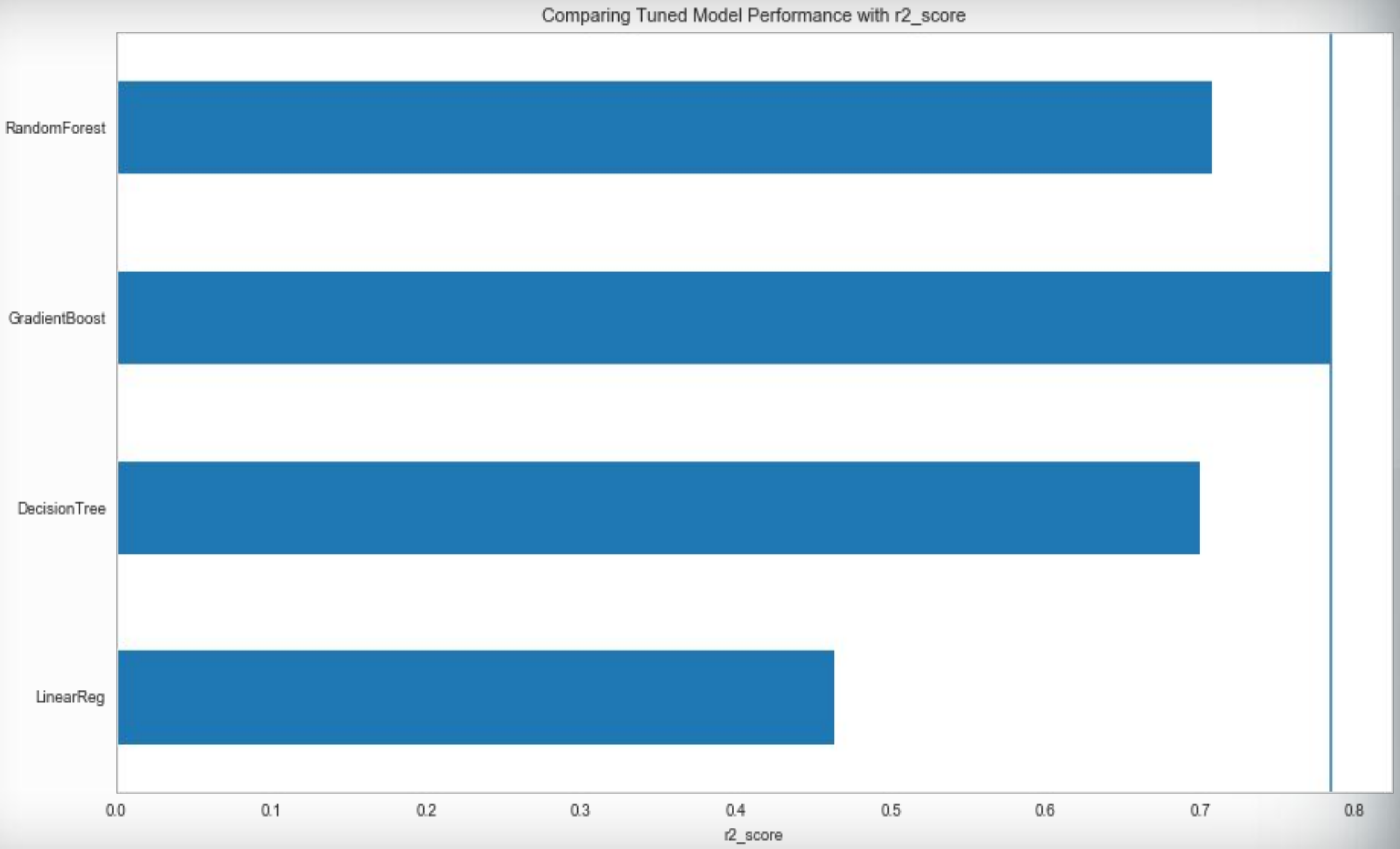


Hyperparameter Tuned Model Performance

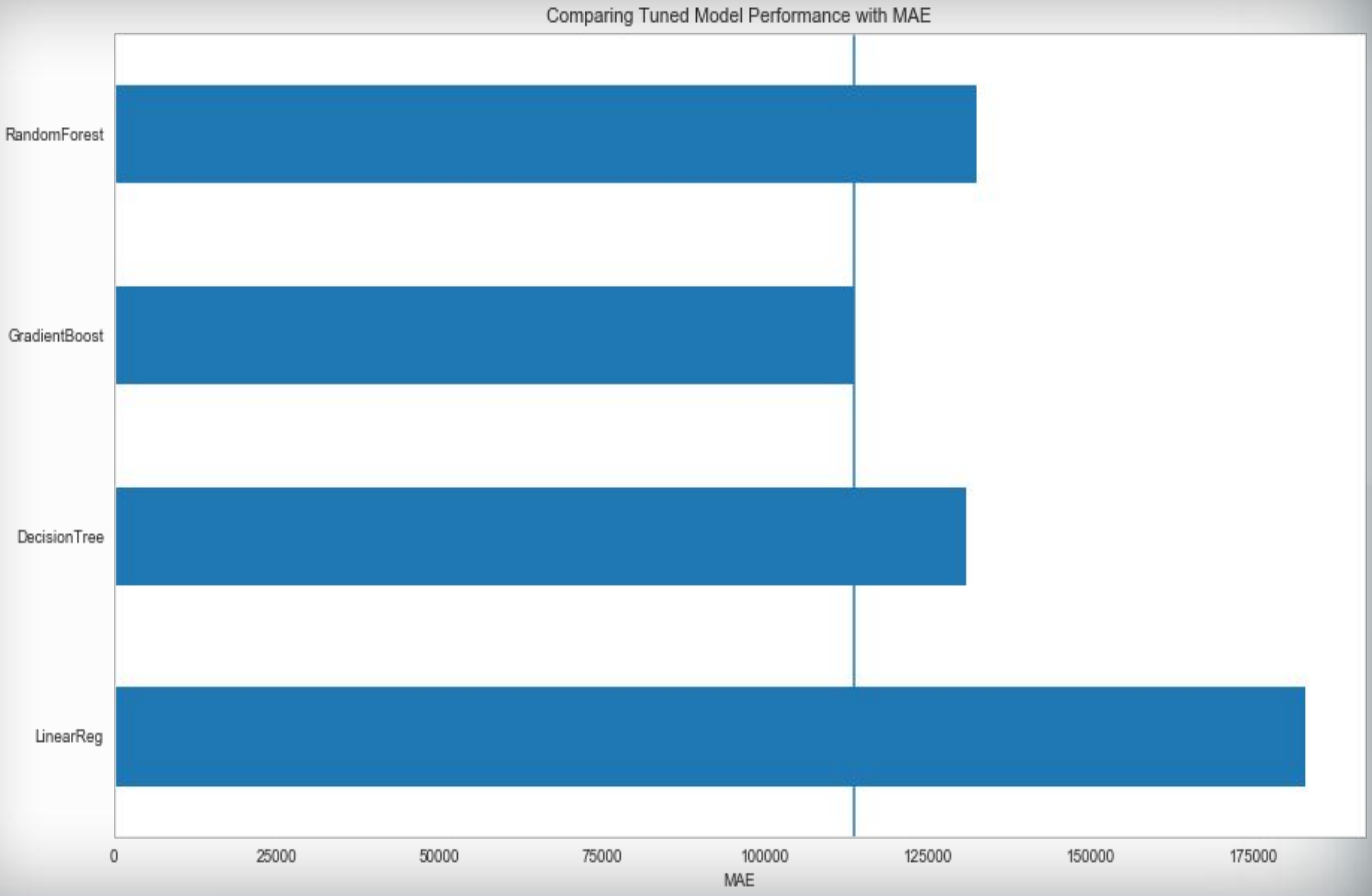
- GridSearchCV
- RandomizedSearchCV



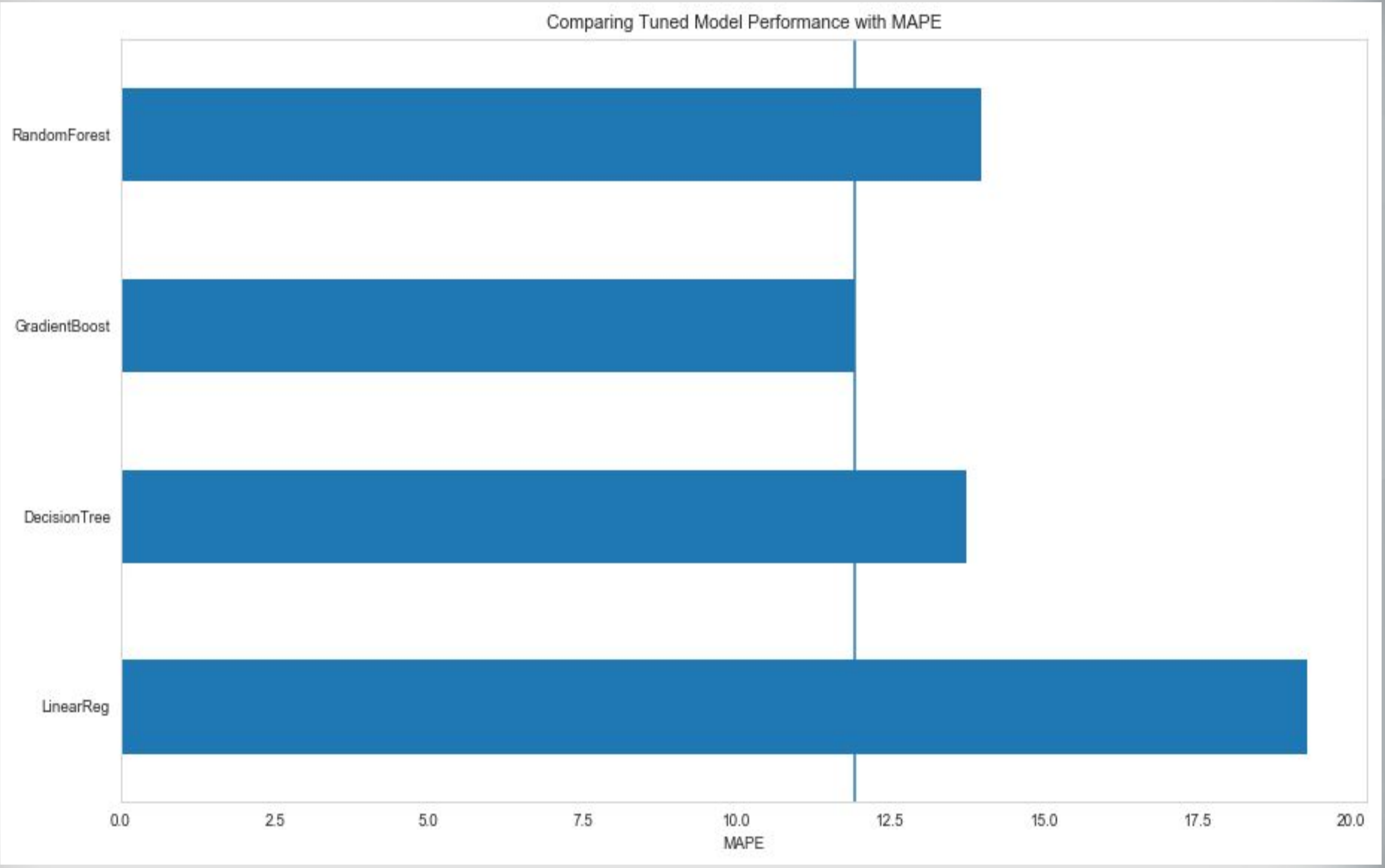
Hyperparameter Tuned Model Performance



Hyperparameter Tuned Model Performance



Hyperparameter Tuned Model Performance



Conclusion

- The Gradient boosting regressor model was better than random guess and it was better performing model compared to other 3 models.
- Many other regression models which were not included in this project can be built for house price prediction.

Appendix

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