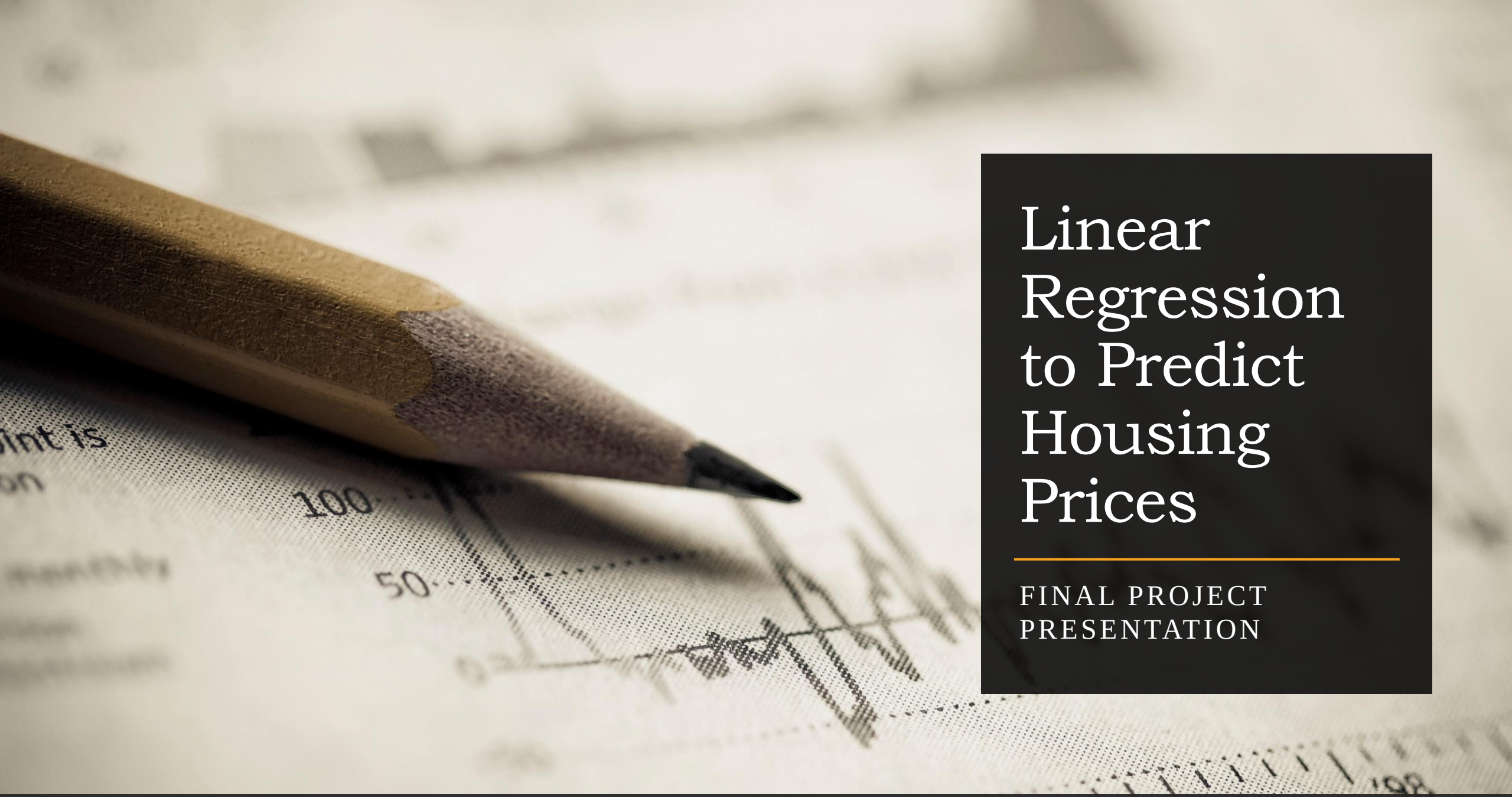


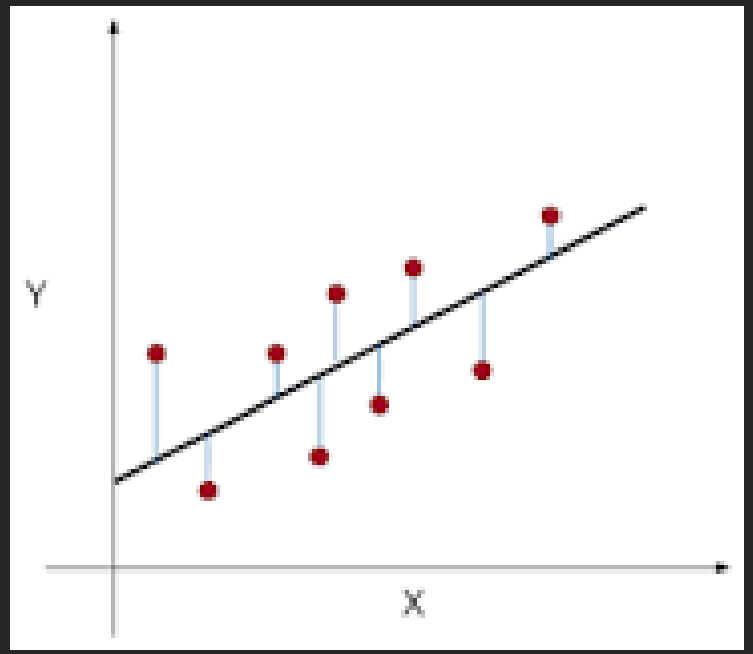
Linear Regression to Predict Housing Prices

FINAL PROJECT PRESENTATION

Introduction

1. To create a regression model that can predict house prices.
2. To show a map of where houses are located using latitude and longitude.
3. To gain a sense of the data, make multiple scatter plots of the price vs. different independent factors.
4. We have to calculate the error for each prediction and how well the model performs on both the training and validation data sets.

Linear Regression



Linear regression models are used to show the relationship between two variables or factors(dependent and independent variables). The factor that is being predicted is called the dependent variable.

From the given data, ‘price’ is an independent variable and the rest are dependent variables.

Why is linear regression used to predict housing prices?

It is a supervised machine learning algorithm in which the projected output is continuous and slope is constant. Instead of classifying data into groups, it is used to forecast values in a continuous range.

Data Exploration

Data exploration is an approach to understand what is in a dataset and the characteristics of the data.

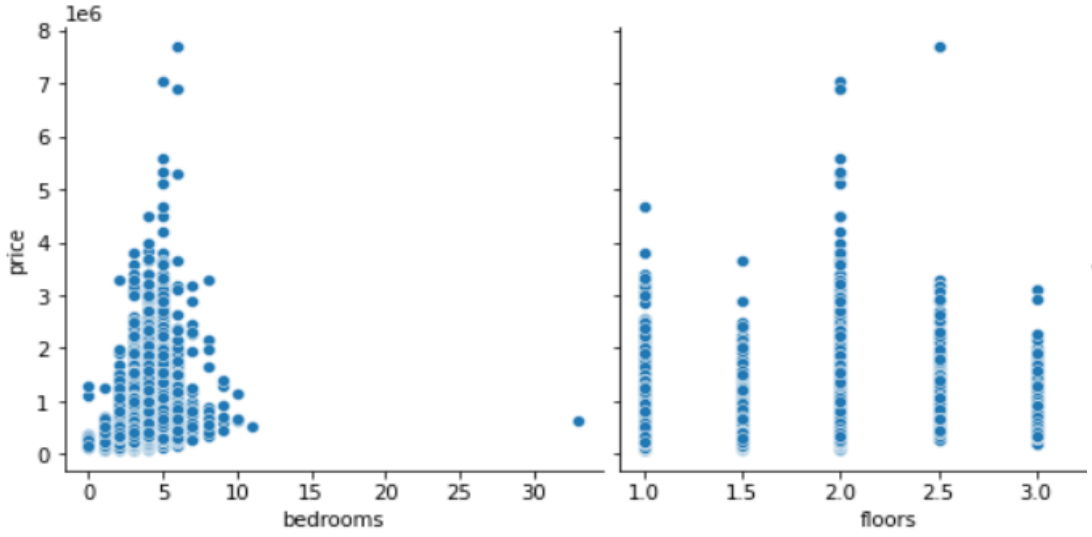
The below table shows the data extracted from Housing Prices dataset 21613,21.

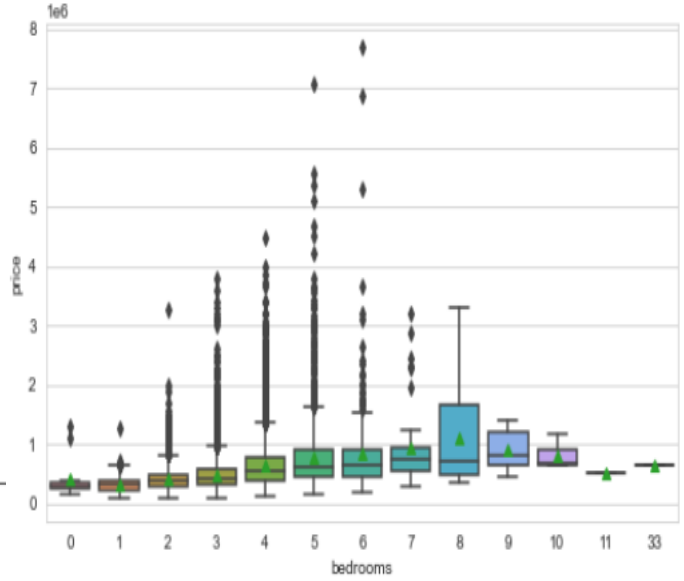
	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0	...	7	1180	0	1955
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0	...	7	2170	400	1951
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0	...	6	770	0	1933
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0	...	7	1050	910	1965
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0	...	8	1680	0	1987

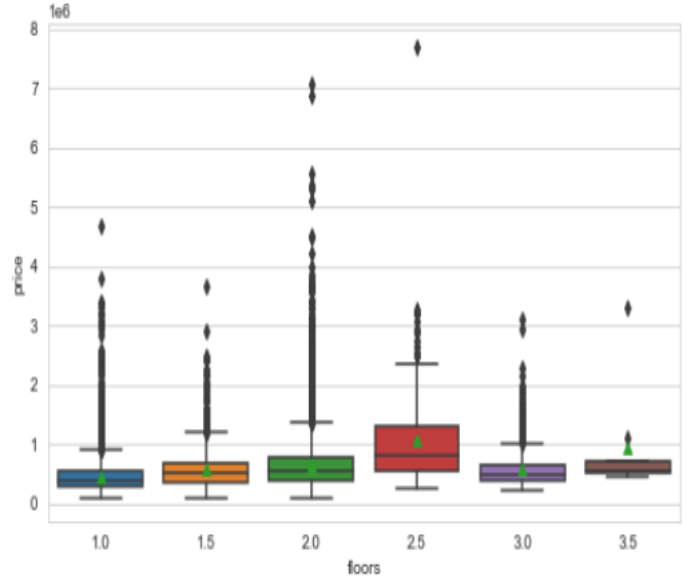
5 rows × 21 columns

- The pairplot() function can be used to plot several pairwise bivariate distributions in a dataset.

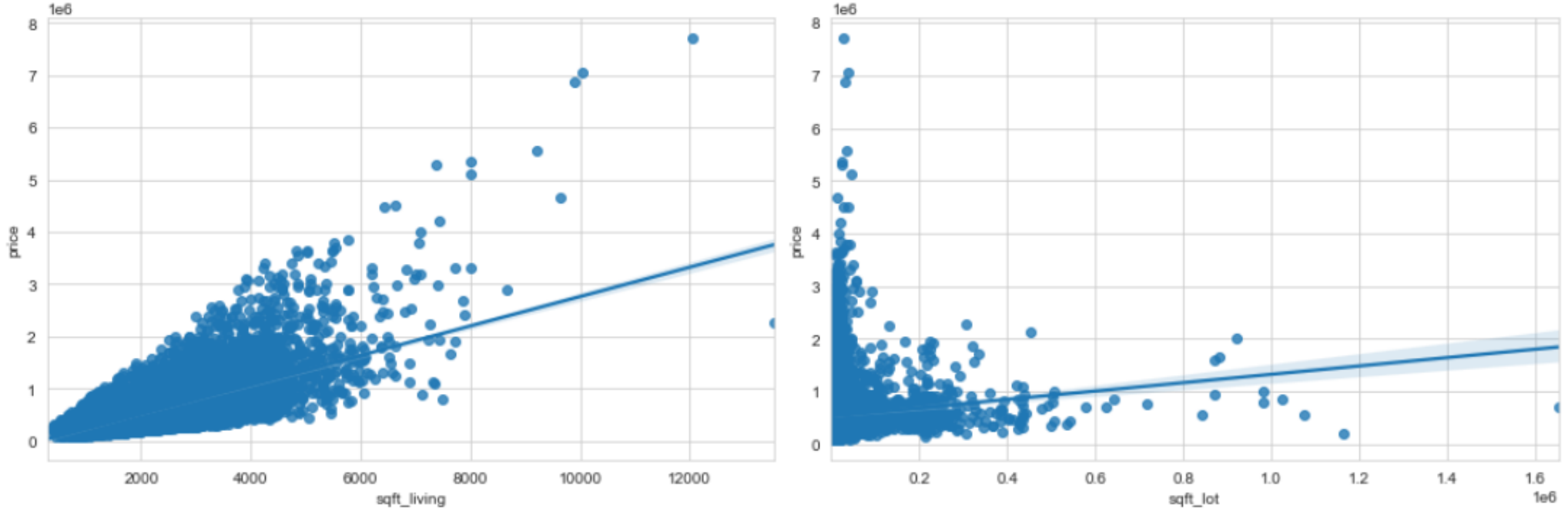
- A box plot is a graphical representation of numerical data's locality, spread, and skewness groups through their quartiles.



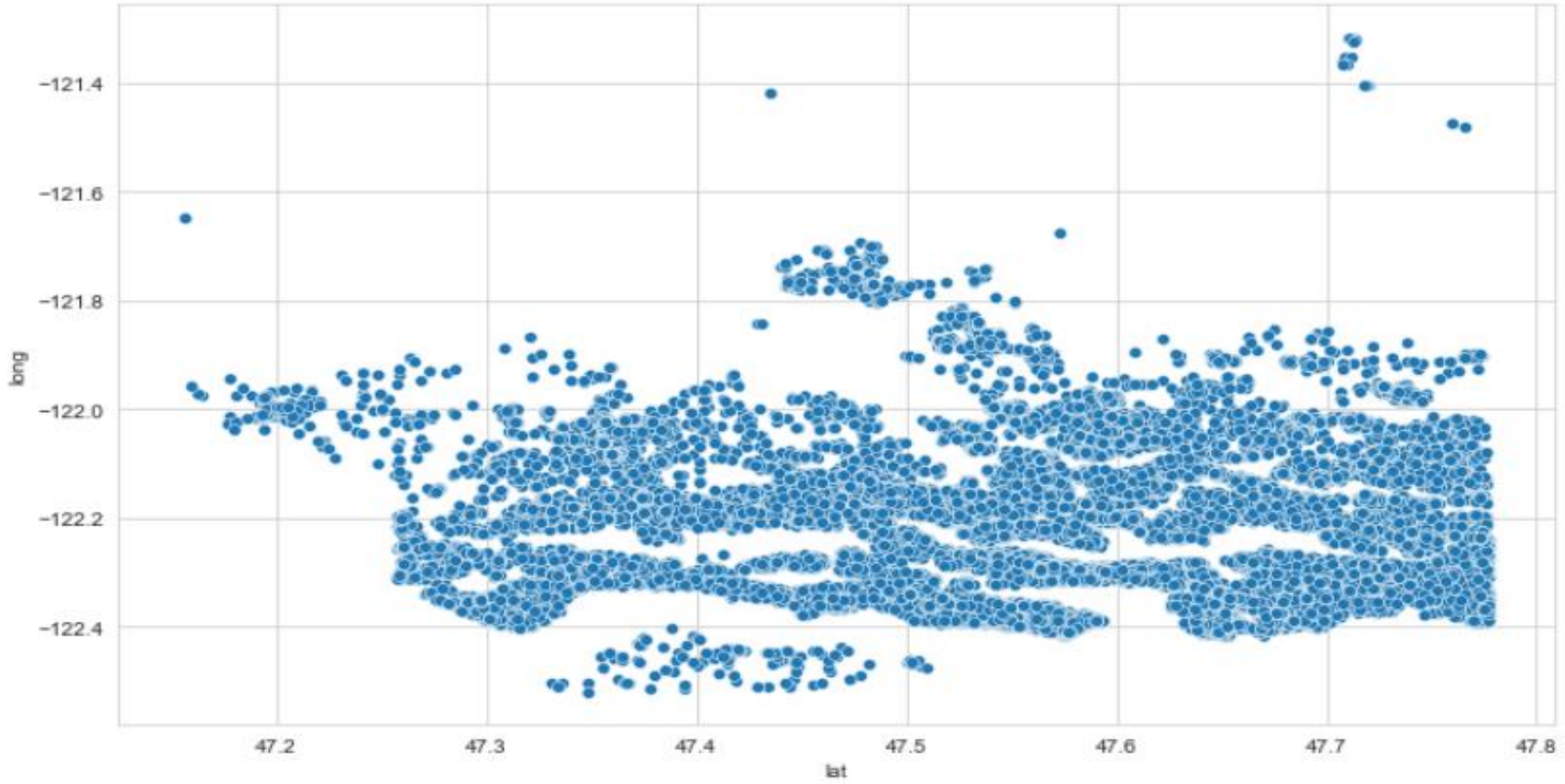


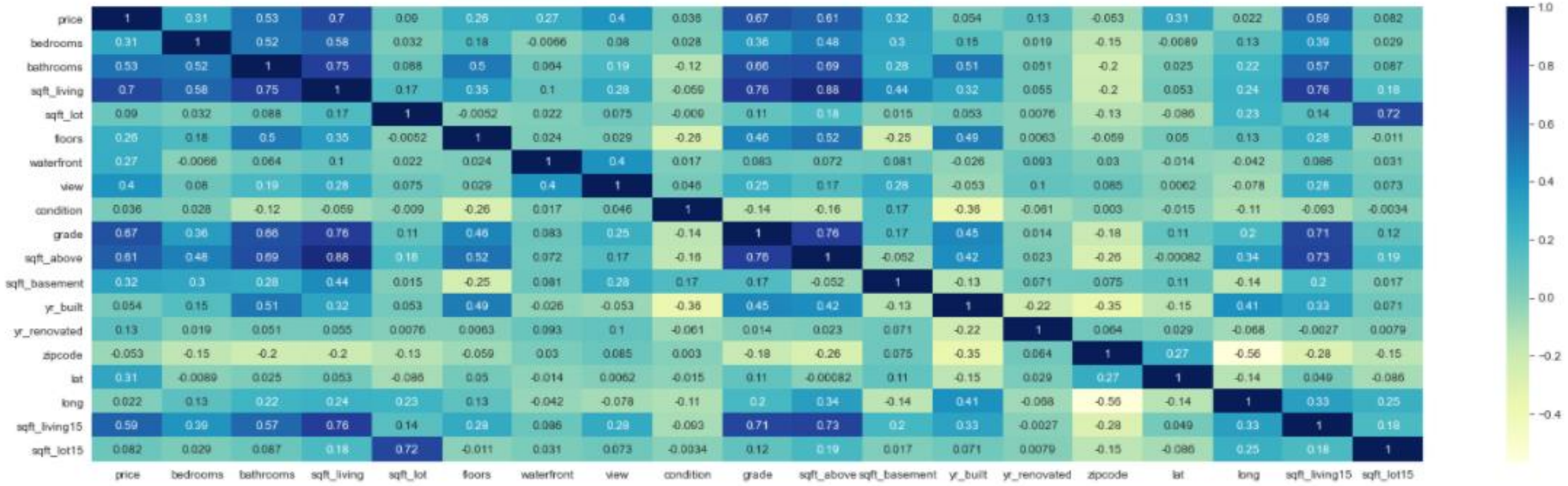


A scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.



Latitude vs Longitude





A heat map is a graphical representation of data where values are depicted by color. Heat maps make it easy to visualize the data.

Splitting the data

```
X = df.drop(['price','id'], axis=1)
y = df['price']
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=23)
```

We must split the data into train and test datasets before we can create the model. The linear regression model will be trained using the train dataset. The test dataset will be used to see if the model becomes overfit and is unable to predict new data not encountered during the training phase.

The Training data is taken as 70% in the mode and the testing data is 30%.

Ordinary Least Squares regression (OLS) is a popular method for calculating the coefficients of linear regression equations that describe the relationship between one or more independent variables and a dependent variable.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.605
Model:	OLS	Adj. R-squared:	0.605
Method:	Least Squares	F-statistic:	3863.
Date:	Sun, 19 Dec 2021	Prob (F-statistic):	0.00
Time:	22:38:47	Log-Likelihood:	-2.0791e+05
No. Observations:	15129	AIC:	4.158e+05
Df Residuals:	15122	BIC:	4.159e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.11e+07	1.68e+06	-24.461	0.000	-4.44e+07	-3.78e+07
waterfront	5.759e+05	2.22e+04	25.909	0.000	5.32e+05	6.19e+05
view	8.294e+04	2688.678	30.849	0.000	7.77e+04	8.82e+04
condition	6.531e+04	2846.738	22.942	0.000	5.97e+04	7.09e+04
grade	1.856e+05	1683.710	110.209	0.000	1.82e+05	1.89e+05
lat	6.261e+05	1.36e+04	46.204	0.000	6e+05	6.53e+05
long	-8.34e+04	1.36e+04	-6.144	0.000	-1.1e+05	-5.68e+04

Omnibus:	11245.210	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	486127.937
Skew:	3.131	Prob(JB):	0.00
Kurtosis:	30.055	Cond. No.	1.21e+05

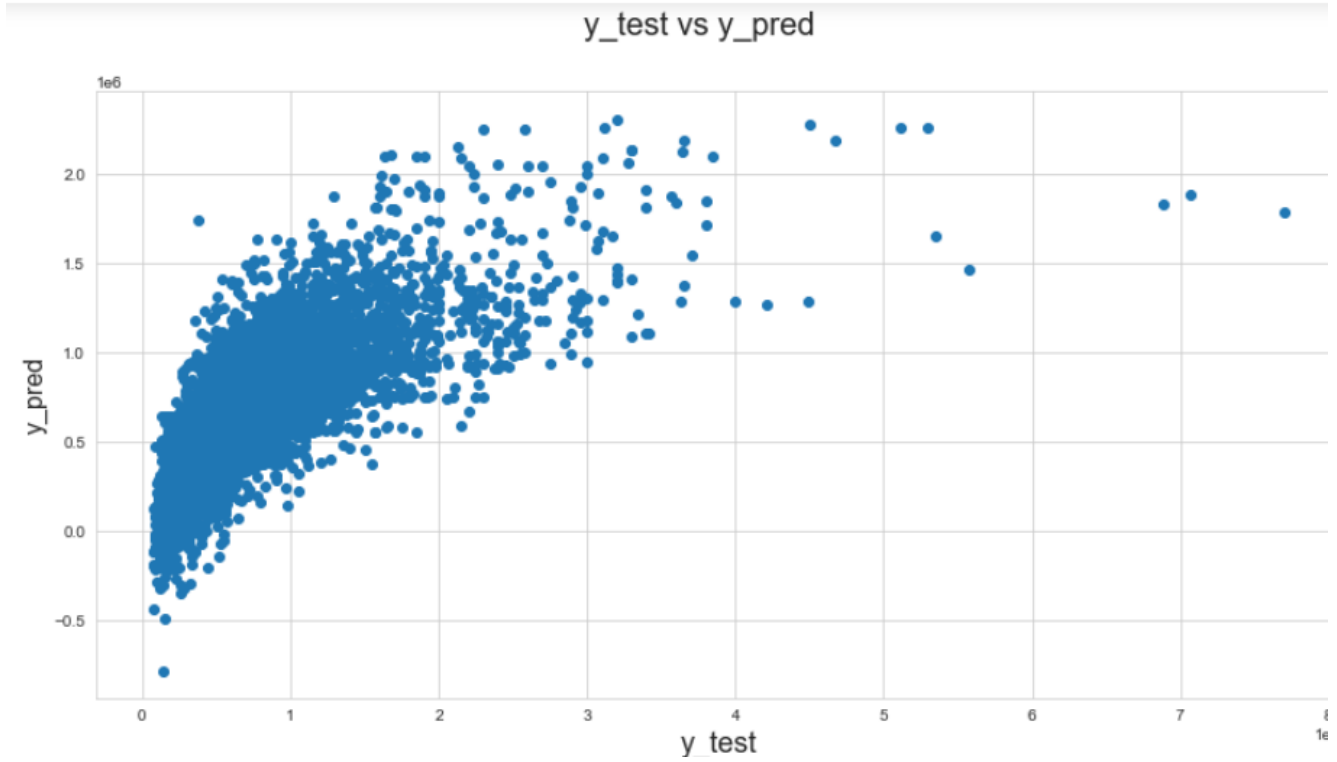
```
In [29]: #to find the r2 score from predicted vs tested.
         from sklearn.metrics import r2_score
         r2_score(Y_test, Y_pred)
```

```
Out[29]: 0.5741846677125284
```

```
In [31]: #plotting the data from tested and predicted data.
         fig = plt.figure(figsize = (14,7))
         plt.scatter(Y_test,Y_pred)
         fig.suptitle('y_test vs y_pred', fontsize=20)
         plt.xlabel('y_test', fontsize=18)
         plt.ylabel('y_pred', fontsize=16)
```

```
Out[31]: Text(0, 0.5, 'y_pred')
```

Graph of tested data vs predicted data



Why is it important to predict Housing Prices

House price predictions are anticipated to assist customers who are planning to buy a home by allowing them to know the price range in the future so that they may properly arrange their finances. House price projections are also useful for property investors who want to know the trajectory of housing prices in a specific area.

Conclusion

- The model's accuracy is fairly good, with a R squared error of 0.605.
- Gradient boosting and Random Forest Regression where the best models so we will use a grid search for both to find the optimal parameters. However, it’ll take more time to process.