Assignment 5

Jett R

5/1/24

In [ ]:
```
import pandas as pd
import numpy as np
```

In [ ]:
```
file = pd.read_csv('C:/Users/jettr/Dropbox (University of Oregon)/23-24/Spring/Geog
```

Task 1

How many houses are in this dataset?

How many features are there for predicting house price?

Are there any null values in this dataset?

Which three variables are best correlated with house price (include correlation coefficients)?

Which three variables are least correlated with house price (include correlation coefficients)?

In [ ]:
```
print('Number of features to prefict house price : ' ,len(file.columns))
print('Number of Houses : ' , len(file['price']))
```

```
Number of features to prefict house price :  8
Number of Houses :  19451
```

In [ ]:
```
file.isna().any()
# There are no null values within the dataset
```

Out[ ]:
```
price          False
bedrooms       False
bathrooms      False
sqft_living    False
sqft_lot       False
yr_built       False
lat            False
long           False
dtype: bool
```

In [ ]:
```
file
```

Out[ ]:

	price	bedrooms	bathrooms	sqft_living	sqft_lot	yr_built	lat	long
0	538000	3	2.25	2570	7242	1951	47.7210	-122.319
1	180000	2	1.00	770	10000	1933	47.7379	-122.233
2	604000	4	3.00	1960	5000	1965	47.5208	-122.393
3	510000	3	2.00	1680	8080	1987	47.6168	-122.045
4	1230000	4	4.50	5420	101930	2001	47.6561	-122.005
...	...	...	...	...	...	...	...	...
19446	475000	3	2.50	1310	1294	2008	47.5773	-122.409
19447	360000	3	2.50	1530	1131	2009	47.6993	-122.346
19448	400000	4	2.50	2310	5813	2014	47.5107	-122.362
19449	400000	3	2.50	1600	2388	2004	47.5345	-122.069
19450	325000	2	0.75	1020	1076	2008	47.5941	-122.299

19451 rows × 8 columns

In [ ]:
```
# Calculate and order the correlations
correlations = file.corr(method='pearson')['price'].sort_values(ascending=False)

# Index 0 is the highest correlation , price, with itself
top_three_correlated = correlations[1:4]
worst_three_correlated = correlations[-3:] # index the 3 worst
print('best correlated : ')
print(top_three_correlated)
print('worst correlated : ')
print(worst_three_correlated)
```

```
best correlated :
sqft_living    0.702296
bathrooms      0.524395
bedrooms       0.315804
Name: price, dtype: float64
worst correlated :
sqft_lot    0.090125
yr_built    0.052453
long        0.020092
Name: price, dtype: float64
```

Task 2

Produce a model to predict house prices. You are welcome to generate new features, scale the data, and split the data into training/testing (i.e. train_test_split) in any way you like. You are also welcome to use the datasets contained in the data folder or other datasets that you find on the internet.

Evaluate your model's accuracy by predicting a test dataset, for example:

On Monday the instructor and TA will provide an unseen set of houses which students will use to repeat their accuracy evaluation. The best models (i.e. lowest RMSE) will win prizes.

We will evaluate the models using a simple mean-squared-error as follows: