

# EDA:

<https://www.kaggle.com/swathiachath/kc-housesales-data>

```
import numpy as np
```

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

```
(sns.set
```

```
import warnings
```

```
from sklearn.model_selection import train_test_split
```

```
from sklearn.metrics import accuracy_score
```

```
('warnings.filterwarnings('ignore
```

```
from sklearn.linear_model import LogisticRegression
```

```
matplotlib inline%
```

```
("2- data=pd.read_csv("f:\\kc_house_data.csv
```

```
()data.head
```

```
3- print('We have {} rows and {} columns'.format(data.shape[0],data.shape[1])
```

Output We have 21613 rows and 21 columns

```
In [8]: #describe the data, which shows numerical data about our data
data.describe()
```

	id	price	bedrooms	bathrooms	sqft_living	sqft_tot	floors	waterfront	view	condition	age
count	21613.000e+04	2.161300e+04	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
mean	4.588332e+08	5.408831e+05	3.378842	2.114757	2879.808738	1.513807e+04	1.484308	0.067542	0.234353	3.489433	7.88
std	2.878884e+08	3.871272e+05	0.938862	0.778163	818.143887	4.142951e+04	0.538868	0.088517	0.788318	0.669743	1.17
min	1.308102e+08	7.500900e+04	0.000000	0.000000	280.000000	5.200000e+02	1.000000	0.000000	0.000000	1.000000	1.00
25%	2.153488e+08	3.218300e+05	3.000000	1.750000	1427.000000	5.049300e+03	1.000000	0.000000	0.000000	3.000000	7.00
50%	2.904920e+08	4.500900e+05	3.000000	2.250000	1810.000000	7.519300e+03	1.500000	0.000000	0.000000	3.000000	7.00
75%	7.308900e+08	8.450900e+05	4.000000	2.500000	2560.000000	1.068900e+04	2.000000	0.000000	0.000000	4.000000	8.00
max	9.808900e+08	7.730800e+06	23.000000	0.000000	13540.000000	1.351258e+08	3.583000	1.000000	4.000000	5.000000	13.00

```
In [10]: #show the data information, show the number of null values in every column
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 21 columns):
#   Column      Non-Null Count  Dtype
---  -
0   id           21613 non-null  int64
1   date        21613 non-null  object
2   price       21613 non-null  float64
3   bedrooms    21613 non-null  int64
4   bathrooms   21613 non-null  float64
```

# EDA:

