HOUSE SALES IN KING COUNTY, USA

Using regression to predict house prices

Data set:
https://www.kaggle.com/harlfoxem/housesalesprediction

ABOUT KING COUNTY

- **King County** is a county located in the US state of Washington.
- Population : 2,188,649
- Average house price : **$634,500**

IMPORTING TOOLS FOR ANALYSIS

In [280]:
```
import pandas as pd
import numpy as np
from scipy import stats
from scipy import stats, special
from sklearn import model_selection, metrics, linear_model, datasets, feature_selection

import matplotlib.pyplot as plt


df = pd.read_csv('kc_house_data.csv.zip')
```

In [281]:
```
df.head()
```

Out[281]:

	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

DATA WRANGLING

```
df['date'] = df['date'].str[0:8]
```

```
df['date'] = pd.to_datetime(df['date'])
```

```
df = df.set_index('id')
```

```
df[df.duplicated()]
```

id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	lat	l

```
df.isnull().values.any()
```

```
False
```

```
df.mean()
```

```
price             540088.141767
bedrooms               3.370842
bathrooms              2.114757
sqft_living         2079.899736
sqft_lot           15106.967566
floors                 1.494309
waterfront             0.007542
view                   0.234303
condition              3.409430
grade                  7.656873
sqft_above          1788.390691
sqft_basement        291.509045
yr_built            1971.005136
yr_renovated          84.402258
lat                   47.560053
long                -122.213896
sqft_living15       1986.552492
sqft_lot15         12768.455652
dtype: float64
```

GETTING THE AVERAGE VALUES FOR A HOUSE IN KING COUNTY.

FINDING RELATIONSHIPS WITHIN THE DATA SET

```
df.cov()
```

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	gra
price	1.347824e+11	105286.276362	148481.495749	2.367154e+08	1.363437e+09	50908.003884	8460.643388	111772.969096	8687.030331	288026.4997
bedrooms	1.052863e+05	0.865015	0.369527	4.925960e+02	1.221324e+03	0.088104	-0.000530	0.056684	0.017232	0.3902
bathrooms	1.484815e+05	0.369527	0.593151	5.338120e+02	2.798944e+03	0.208211	0.004247	0.110800	-0.062638	0.6020
sqft_living	2.367154e+08	492.596040	533.811988	8.435337e+05	6.574684e+06	175.540402	8.249461	200.314304	-35.114601	823.4076
sqft_lot	1.363437e+09	1221.324216	2798.943628	6.574684e+06	1.715659e+09	-116.328567	77.418670	2371.393311	-241.461641	5531.9969
floors	5.090800e+04	0.088104	0.208211	1.755404e+02	-1.163286e+02	0.291588	0.001107	0.012184	-0.092686	0.2908
waterfront	8.460643e+03	-0.000530	0.004247	8.249461e+00	7.741867e+01	0.001107	0.007485	0.026643	0.000938	0.0084
view	1.117730e+05	0.056684	0.110800	2.003143e+02	2.371393e+03	0.012184	0.026643	0.587243	0.022934	0.2263
condition	8.687030e+03	0.017232	-0.062638	-3.511460e+01	-2.414616e+02	-0.092686	0.000938	0.022934	0.423467	-0.1106
grade	2.880265e+05	0.390254	0.602005	8.234077e+02	5.531997e+03	0.290824	0.008418	0.226383	-0.110664	1.3817
sqft_above	1.841014e+08	367.836092	437.087555	6.666978e+05	6.294462e+06	234.260251	5.163720	106.387032	-85.257275	735.8054
sqft_basement	5.261406e+07	124.759948	96.724433	1.768358e+05	2.802218e+05	-58.719850	3.085741	93.927272	50.142673	87.6022
yr_built	5.824484e+05	4.212004	11.447333	8.580238e+03	6.458085e+04	7.761250	-0.066483	-1.202897	-6.908312	15.4324
yr_renovated	1.864486e+07	7.038678	15.696537	2.042442e+04	1.271708e+05	1.374814	3.227949	31.987181	-15.844882	6.8058
lat	1.561742e+04	-0.001151	0.002622	6.685035e+00	-4.917661e+02	0.003712	-0.000171	0.000654	-0.001347	0.0185
long	1.118118e+03	0.016958	0.024191	3.107108e+01	1.338837e+03	0.009538	-0.000511	-0.008461	-0.009760	0.0328
sqft_living15	1.472964e+08	249.651804	300.161076	4.761601e+05	4.105319e+06	103.586570	5.127103	147.294289	-41.400888	574.5907
sqft_lot15	8.264591e+08	742.644640	1833.182173	4.596302e+06	8.126540e+08	-166.152367	72.529786	1518.526494	-60.509350	3827.2537

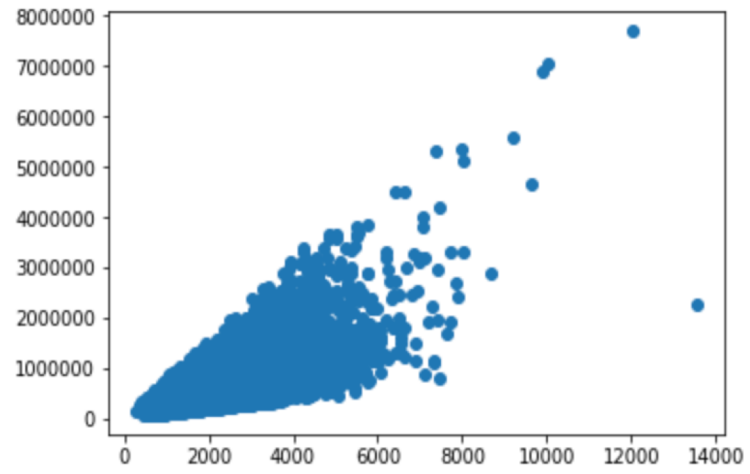
DETERMINED PRIME FACTORS THAT AFFECT THE HOUSE PRICE

```
x = df.corr()
x
```

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built
price	1.000000	0.308350	0.525138	0.702035	0.089661	0.256794	0.266369	0.397293	0.036362	0.667434	0.605567	0.323816	0.054012
bedrooms	0.308350	1.000000	0.515884	0.576671	0.031703	0.175429	-0.006582	0.079532	0.028472	0.356967	0.477600	0.303093	0.154178
bathrooms	0.525138	0.515884	1.000000	0.754665	0.087740	0.500653	0.063744	0.187737	-0.124982	0.664983	0.685342	0.283770	0.506019
sqft_living	0.702035	0.576671	0.754665	1.000000	0.172826	0.353949	0.103818	0.284611	-0.058753	0.762704	0.876597	0.435043	0.318049
sqft_lot	0.089661	0.031703	0.087740	0.172826	1.000000	-0.005201	0.021604	0.074710	-0.008958	0.113621	0.183512	0.015286	0.053080
floors	0.256794	0.175429	0.500653	0.353949	-0.005201	1.000000	0.023698	0.029444	-0.263768	0.458183	0.523885	-0.245705	0.489319
waterfront	0.266369	-0.006582	0.063744	0.103818	0.021604	0.023698	1.000000	0.401857	0.016653	0.082775	0.072075	0.080588	-0.026161
view	0.397293	0.079532	0.187737	0.284611	0.074710	0.029444	0.401857	1.000000	0.045990	0.251321	0.167649	0.276947	-0.053440
condition	0.036362	0.028472	-0.124982	-0.058753	-0.008958	-0.263768	0.016653	0.045990	1.000000	-0.144674	-0.158214	0.174105	-0.361417
grade	0.667434	0.356967	0.664983	0.762704	0.113621	0.458183	0.082775	0.251321	-0.144674	1.000000	0.755923	0.168392	0.446963
sqft_above	0.605567	0.477600	0.685342	0.876597	0.183512	0.523885	0.072075	0.167649	-0.158214	0.755923	1.000000	-0.051943	0.423898
_basement	0.323816	0.303093	0.283770	0.435043	0.015286	-0.245705	0.080588	0.276947	0.174105	0.168392	-0.051943	1.000000	-0.133124
yr_built	0.054012	0.154178	0.506019	0.318049	0.053080	0.489319	-0.026161	-0.053440	-0.361417	0.446963	0.423898	-0.133124	1.000000
_renovated	0.126434	0.018841	0.050739	0.055363	0.007644	0.006338	0.092885	0.103917	-0.060618	0.014414	0.023285	0.071323	-0.224874
lat	0.307003	-0.008931	0.024573	0.052529	-0.085683	0.049614	-0.014274	0.006157	-0.014941	0.114084	-0.000816	0.110538	-0.148122
long	0.021626	0.129473	0.223042	0.240223	0.229521	0.125419	-0.041910	-0.078400	-0.106500	0.198372	0.343803	-0.144765	0.409356
qft_living15	0.585379	0.391638	0.568634	0.756420	0.144608	0.279885	0.086463	0.280439	-0.092824	0.713202	0.731870	0.200355	0.326229
sqft_lot15	0.082447	0.029244	0.087175	0.183286	0.718557	-0.011269	0.030703	0.072575	-0.003406	0.119248	0.194050	0.017276	0.070958

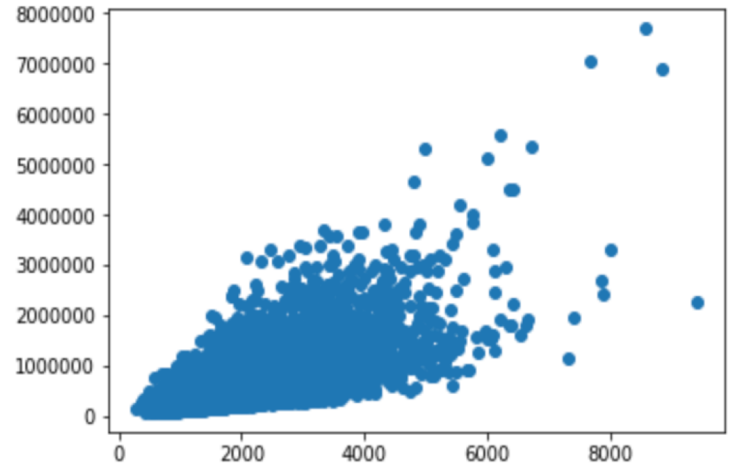
```
plt.scatter(x,y)
```

```
<matplotlib.collections.PathCollection at 0x1a17daf8d0>
```




```
plt.scatter(x1,y)
```

```
<matplotlib.collections.PathCollection at 0x10b932358>
```

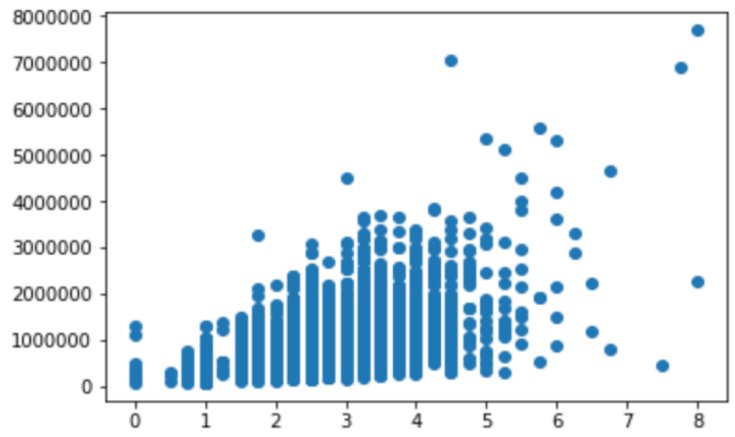


Graph 1 : Sqft Living vs. Price

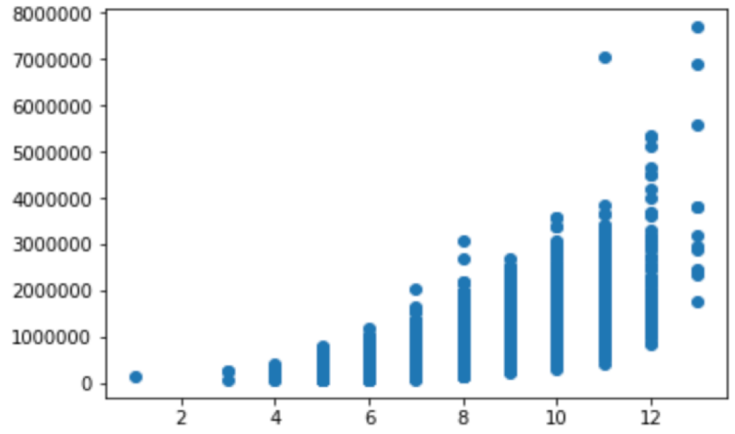
Graph 2 : Sqft Above vs. Price

Both scatter plot graphs show a linear relationship, proving that they have a big impact on the price of house.

```
In [18]: plt.scatter(x2,y)
Out[18]: <matplotlib.collections.PathCollection at 0x10b8544a8>
```




```
In [19]: plt.scatter(x3,y)
Out[19]: <matplotlib.collections.PathCollection at 0x1a17d73d30>
```



Graph 1: Bathroom vs. Price

Graph 2: Grade vs. Price

Both graphs show a strong linear relation, meaning they also impact the house price.

USING TEST_TRAIN_SPLIT METHOD FOR DATA VALIDATION

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

```
X5_train, X5_test, y_train , y_test =  train_test_split(X5,y,test_size=0.2)
```

```
len(X5_train)
```

```
17290
```

```
len(X5_test)
```

```
4323
```

* X5 = df[['sqft_living','sqft_above','bathrooms','grade','sqft_living15','yr_renovated','lat', 'long','view','waterfront']]
y = df['price']

TESTING THE MODEL

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
```

```
model.fit(X5_train,y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

```
model.predict(X5_test)
```

```
array([581013.80884069, 146497.89796319,  58385.60739829, ...,
       467155.85153744, 481234.32233165, 318734.63116586])
```

```
model.score(X5_test,y_test)
```

```
0.66958667294963
```

```
model.fit(df[['sqft_living','sqft_above','bathrooms','grade','sqft_living15','yr_renovated','lat', 'long','view','water
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

CONCLUSION

- The model was able to predict the prices with 66% accuracy.
- The main factors affecting the price are : Square footage , bathrooms , and the grade of house.