Linear Regression with Python

This is mostly just code for reference. Please watch the video lecture for more info behind all of this code.

Your neighbor is a real estate agent and wants some help predicting housing prices for regions in the USA. It would be great if you could somehow create a model for her that allows her to put in a few features of a house and returns back an estimate of what the house would sell for.

She has asked you if you could help her out with your new data science skills. You say yes, and decide that Linear Regression might be a good path to solve this problem!

Your neighbor then gives you some information about a bunch of houses in regions of the United States,it is all in the data set: USA_Housing.csv.

The data contains the following columns:

- 'Avg. Area Income': Avg. Income of residents of the city house is located in.
- 'Avg. Area House Age': Avg Age of Houses in same city
- 'Avg. Area Number of Rooms': Avg Number of Rooms for Houses in same city
- 'Avg. Area Number of Bedrooms': Avg Number of Bedrooms for Houses in same city
- 'Area Population': Population of city house is located in
- 'Price': Price that the house sold at
- 'Address': Address for the house

Let's get started!

Check out the data

We've been able to get some data from your neighbor for housing prices as a csv set, let's get our environment ready with the libraries we'll need and then import the data!

Import Libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Check out the Data

In [2]:
```
USAhousing = pd.read_csv('USA_Housing.csv')
```

In [3]:
```
USAhousing.head()
```

Out[3]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michae 674\nLaura 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnso Suite 079\n Kathleen, C
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizab Stravenue\r WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barne 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Rayr AE 09386

In [5]:
```
USAhousing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 7 columns):
Avg. Area Income                5000 non-null float64
Avg. Area House Age             5000 non-null float64
Avg. Area Number of Rooms       5000 non-null float64
Avg. Area Number of Bedrooms    5000 non-null float64
Area Population                 5000 non-null float64
Price                           5000 non-null float64
Address                         5000 non-null object
dtypes: float64(6), object(1)
memory usage: 273.5+ KB
```

In [6]:
```
USAhousing.describe()
```

Out[6]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Pri
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+

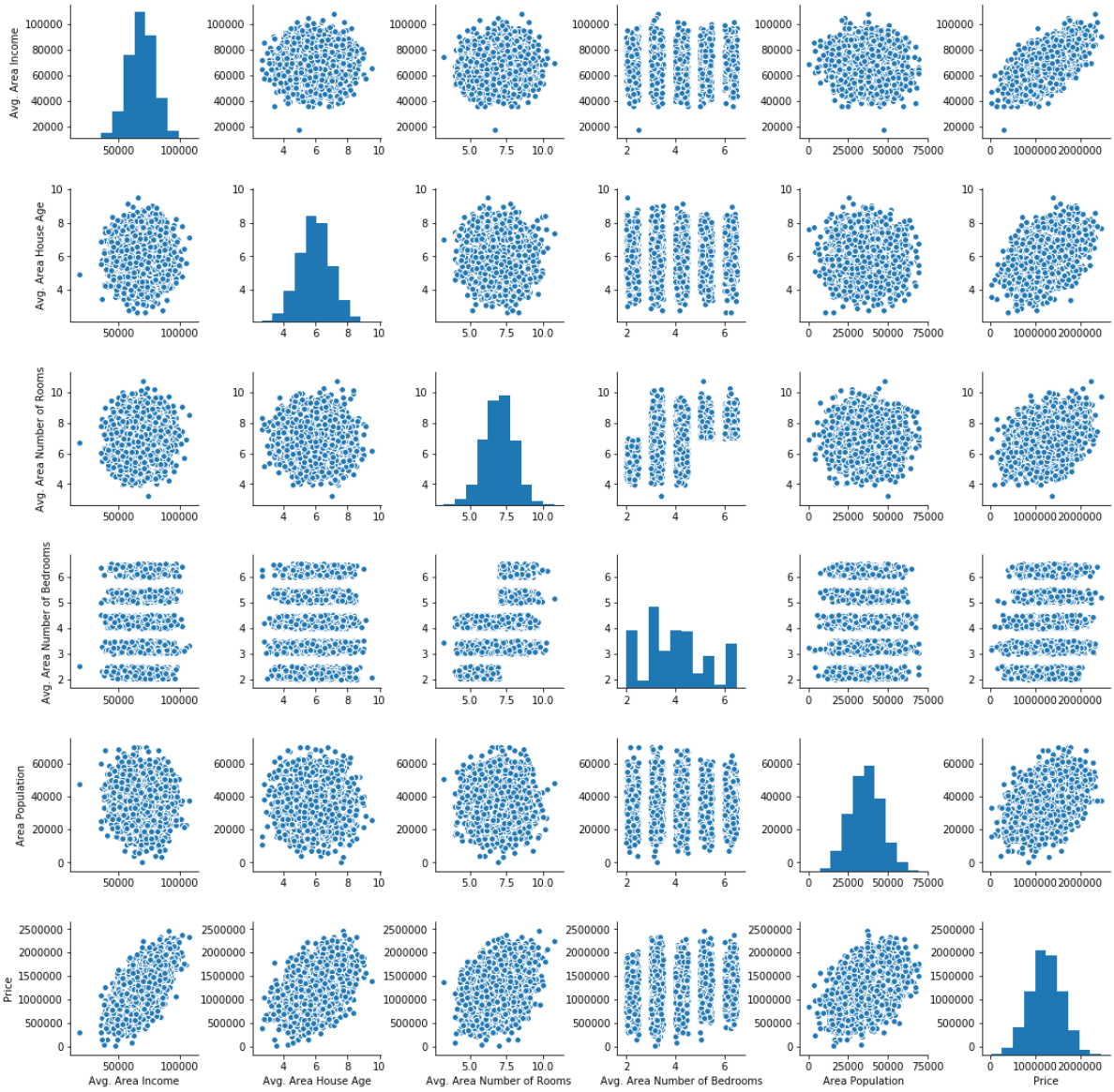
In [7]:
```
USAhousing.columns
```

Out[7]:
```
Index(['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Room
s',
       'Avg. Area Number of Bedrooms', 'Area Population', 'Price', 'Addres
s'],
      dtype='object')
```

Exploratory Data Analysis

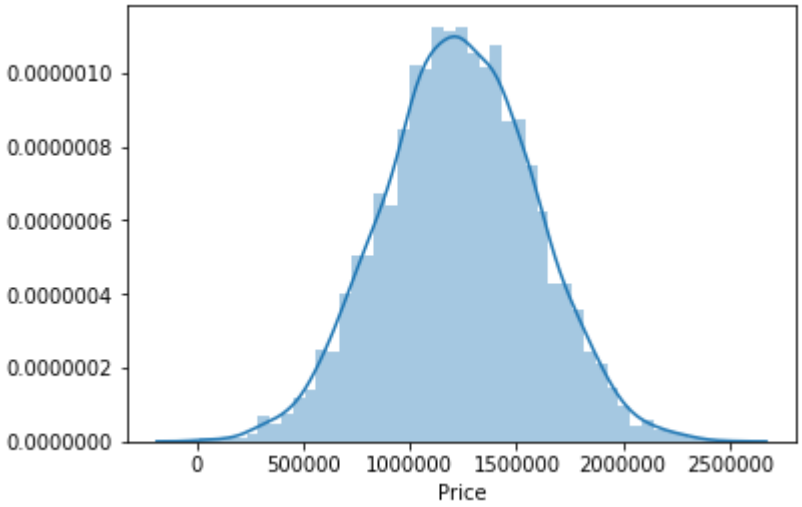
In [8]:
```
sns.pairplot(USAhousing)
```

Out[8]: <seaborn.axisgrid.PairGrid at 0xbfba438>



In [9]:
```
sns.distplot(USAhousing['Price'])
```

Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0xee20630>



In [12]:
```
sns.heatmap(USAhousing.corr(), annot=True, cmap='viridis')
```

Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0xe790da0>

Training a Linear Regression Model

Let's now begin to train out regression model! We will need to first split up our data into an X array that contains the features to train on, and a y array with the target variable, in this case the Price column. We will toss out the Address column because it only has text info that the linear regression model can't use.

Training & Testing Data Set

In [13]:
```
X = USAhousing[['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number o
f Rooms',
       'Avg. Area Number of Bedrooms', 'Area Population']]
y = USAhousing['Price']
```

In [14]:
```
from sklearn.model_selection import train_test_split
```

In [15]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, rando
m_state=101)
```

Model Initialisation & Training

In [16]:
```
from sklearn.linear_model import LinearRegression
```

In [17]:
```
lm = LinearRegression()
```

In [25]:
```
lm.fit(X_train, y_train)
```

Out[25]:
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

Model Evaluation

Using 'intercept' & 'coefficients'

In [26]:
```
print(lm.intercept_)
```

```
-2641372.6673
```

In [27]:
```
coeff_df = pd.DataFrame(lm.coef_, index=X.columns, columns=['Coefficients'] )
```

In [29]:
```
coeff_df
```

Out[29]:

	Coefficients
Avg. Area Income	21.617635
Avg. Area House Age	165221.119872
Avg. Area Number of Rooms	121405.376596
Avg. Area Number of Bedrooms	1318.718783
Area Population	15.225196

Interpreting the coefficients:

- Holding all other features fixed, a 1 unit increase in **Avg. Area Income** is associated with an **increase of $21.52** .
- Holding all other features fixed, a 1 unit increase in **Avg. Area House Age** is associated with an **increase of $164883.28** .
- Holding all other features fixed, a 1 unit increase in **Avg. Area Number of Rooms** is associated with an **increase of $122368.67** .
- Holding all other features fixed, a 1 unit increase in **Avg. Area Number of Bedrooms** is associated with an **increase of $2233.80** .
- Holding all other features fixed, a 1 unit increase in **Area Population** is associated with an **increase of $15.15** .

Does this make sense? Probably not because this is a made up this data.