In [30]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set()

import warnings
warnings.filterwarnings("ignore")

from sklearn.linear_model import LinearRegression
```

In [5]:
```
raw_data = pd.read_csv("USA_Housing.csv")
raw_data.head()
```

Out[5]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386

In [16]:
```
raw_data.describe()
```

Out[16]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+05
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06

In [17]:
```
raw_data.isnull().sum()
```

Out[17]:
```
Avg. Area Income              0
Avg. Area House Age           0
Avg. Area Number of Rooms     0
Avg. Area Number of Bedrooms  0
Area Population               0
Price                         0
Address                       0
dtype: int64
```

In [20]:
```
data = raw_data.drop("Address", axis=1)
data.describe()
```

Out[20]:

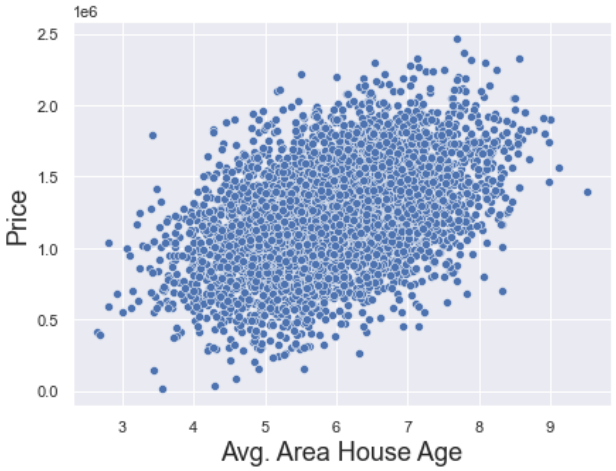
	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+05
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06

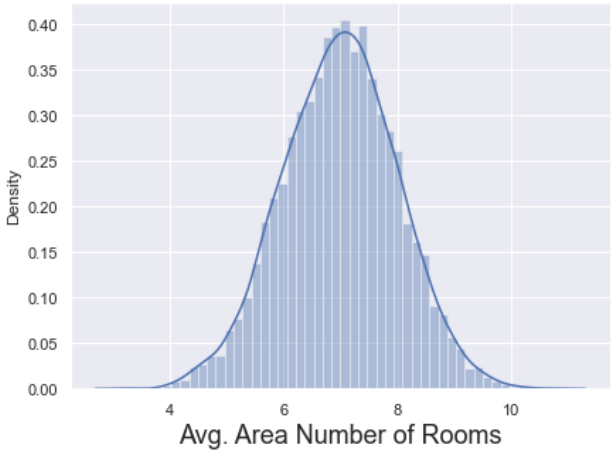
In [160…
```
target = data.Price
inputs = data.drop("Price",axis=1)
```

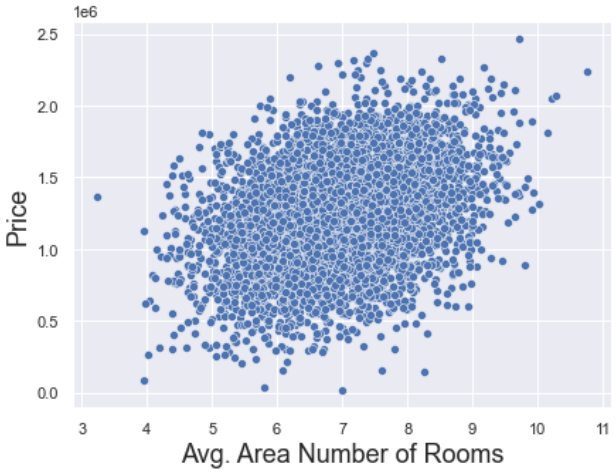
In [163…
```
def scatter_distplot(col):
    fig, axes = plt.subplots(1,2,figsize=(15,5))

    scatter_plot = sns.scatterplot(x=inputs[col], y=target, ax=axes[1])
    scatter_plot.set_xlabel(col, fontsize=18)
    scatter_plot.set_ylabel("Price", fontsize=18)

    dist_plot = sns.distplot(inputs[col], ax=axes[0])
    dist_plot.set_xlabel(col, fontsize=18)
```

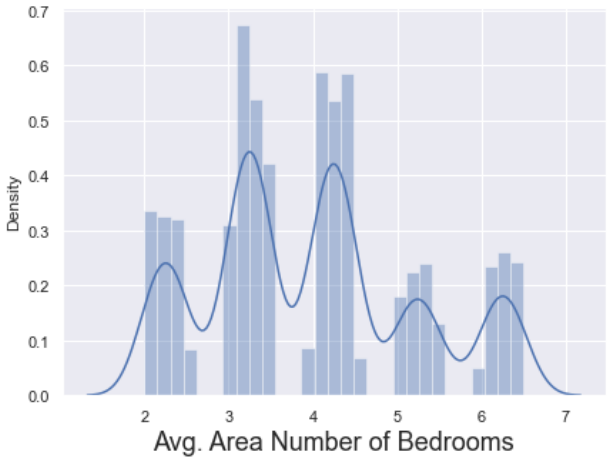
In [164…

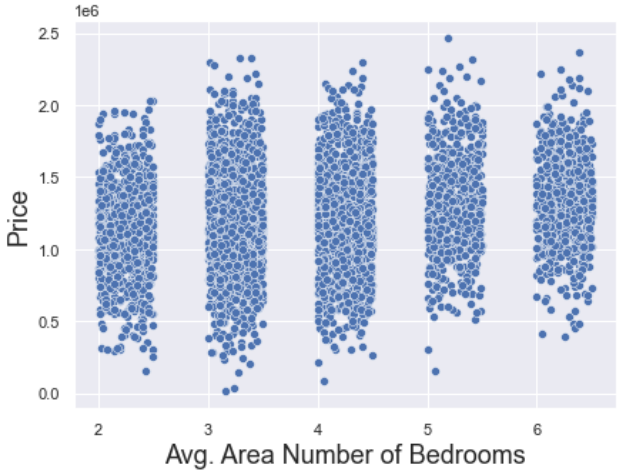
```
for col in inputs.columns:
    scatter_distplot(col)
```

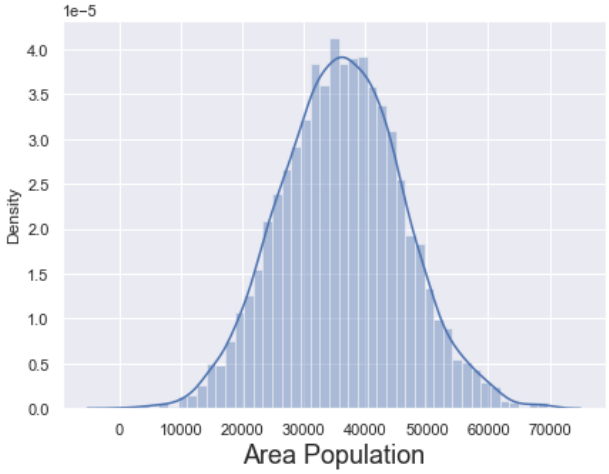












In [166…
```
from statsmodels.stats.outliers_influence import variance_inflation_factor

vif = pd.DataFrame()
vif["features"] = inputs.columns
vif["VIF"] = [variance_inflation_factor(inputs.values, i) for i in range(inputs.shape[1
vif
```

Out[166…

	features	VIF
0	Avg. Area Income	29.650899
1	Avg. Area House Age	27.447775
2	Avg. Area Number of Rooms	45.257291
3	Avg. Area Number of Bedrooms	14.537873
4	Area Population	12.825450

In [167…
```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(inputs)
```

Out[167…
```
StandardScaler()
```

In [168…
```
inputs_scaled = scaler.transform(inputs)
```

In [174…
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(inputs_scaled, target, test_size=0.
```

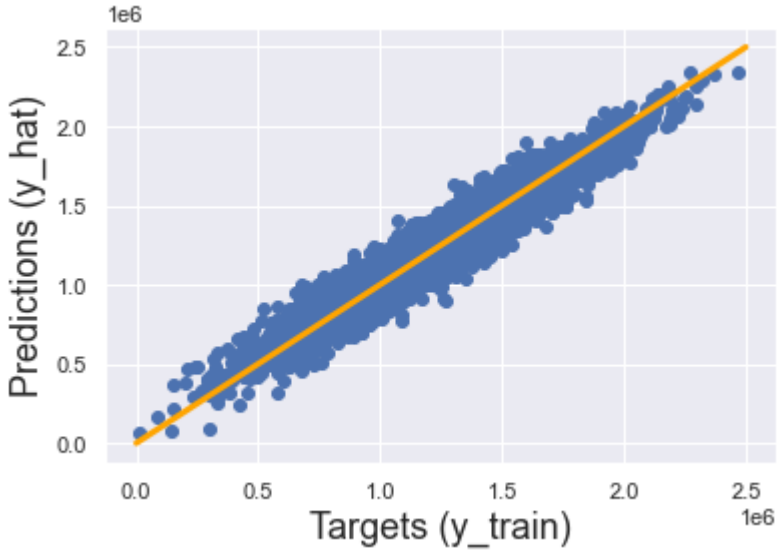
In [175…
```
reg = LinearRegression()
reg.fit(x_train, y_train)
```

Out[175…
```
LinearRegression()
```

In [176…
```
y_hat = reg.predict(x_train)
```

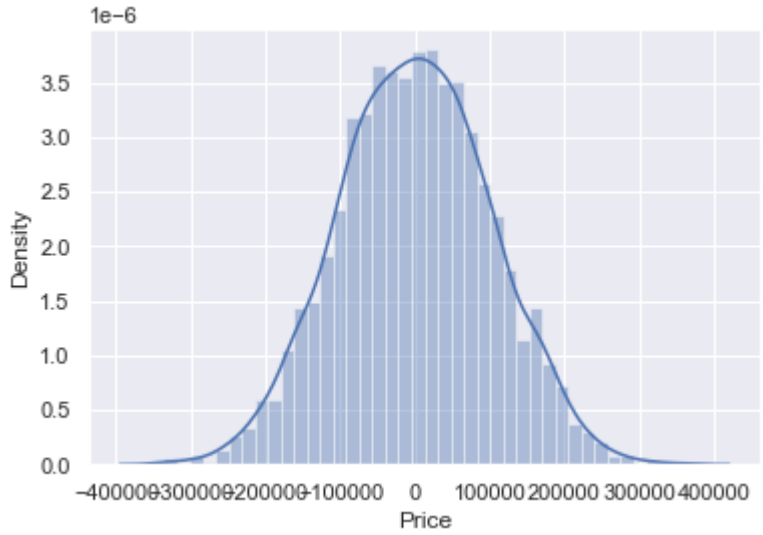
In [185…
```
plt.scatter(y_train, y_hat)
plt.xlabel("Targets (y_train)", fontsize=18)
plt.ylabel("Predictions (y_hat)", fontsize=18)

x = np.linspace(0,2.5e6)
y = x
plt.plot(x,y,c="orange",lw=3)
```

Out[185…
```
[<matplotlib.lines.Line2D at 0x26701a6b310>]
```



In [188…
```
residuals_train = y_train - y_hat
sns.distplot(residuals_train)
residuals_train.describe()
```

Out[188…
```
count    4.000000e+03
mean    -5.312904e-11
std      1.015767e+05
min     -3.370071e+05
25%     -6.991187e+04
50%     -1.058613e+02
75%      6.895613e+04
max      3.624884e+05
Name: Price, dtype: float64
```



In [189…
```
reg.score(x_train, y_train)
```

Out[189…
```
0.9181859079129733
```

In [190…
```
reg.intercept_
```

Out[190…
```
1232374.526139742
```

In [191…
```
reg.coef_
```

Out[191…
```
array([230342.10551988, 164805.29545772, 120130.82515573,   2723.03223192,
       151552.41007222])
```

In [195…
```
reg_summary = pd.DataFrame(columns=["features"], data=inputs.columns)
reg_summary["weights"] = reg.coef_
reg_summary
```

Out[195…

	features	weights
0	Avg. Area Income	230342.105520
1	Avg. Area House Age	164805.295458
2	Avg. Area Number of Rooms	120130.825156
3	Avg. Area Number of Bedrooms	2723.032232
4	Area Population	151552.410072

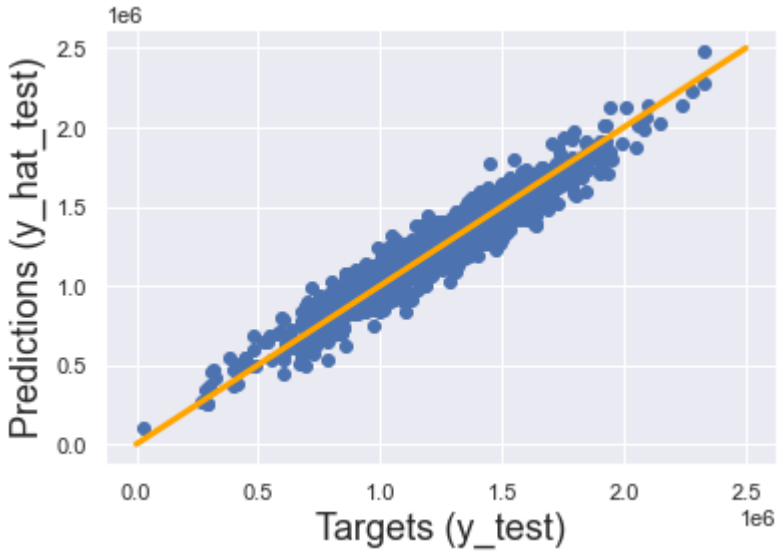
In [196…
```
y_hat_test = reg.predict(x_test)
```

In [197…
```
plt.scatter(y_test, y_hat_test)
plt.xlabel("Targets (y_test)", fontsize=18)
plt.ylabel("Predictions (y_hat_test)", fontsize=18)
```

```
x = np.linspace(0,2.5e6)
y = x
plt.plot(x,y,c="orange",lw=3)
```

Out[197…

```
[<matplotlib.lines.Line2D at 0x2677f0230d0>]
```



In [198…

```
reg.score(x_test, y_test)
```

Out[198…

```
0.9172058023346101
```