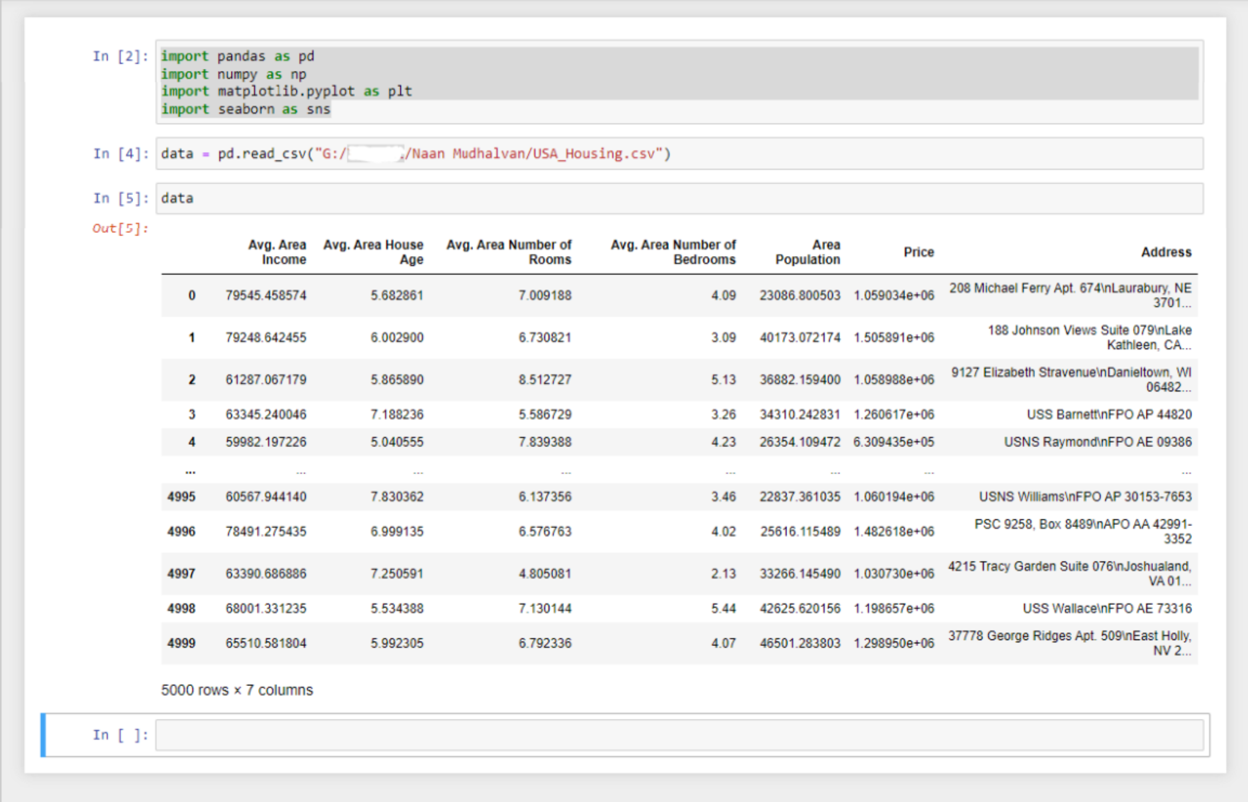
Loading and Preprocessing the Dataset

Loading the Dataset

- **Using pandas.read_csv() function, we read the USA_Housing dataset.**



```
In [2]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns

In [4]: data = pd.read_csv("G:/[redacted]/Naan Mudhalvan/USA_Housing.csv")

In [5]: data
Out[5]:
```

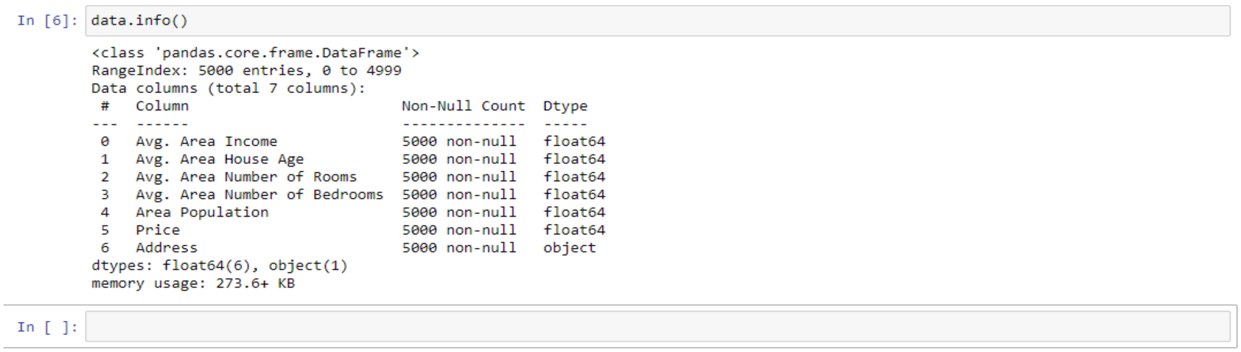
	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 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
...	...	...	...	...	...	...	...
4995	60567.944140	7.830362	6.137356	3.46	22837.361035	1.060194e+06	USNS Williams\nFPO AP 30153-7653
4996	78491.275435	6.999135	6.576763	4.02	25616.115489	1.482618e+06	PSC 9258, Box 8489\nAPO AA 42991-3352
4997	63390.686886	7.250591	4.805081	2.13	33266.145490	1.030730e+06	4215 Tracy Garden Suite 076\nJoshualand, VA 01...
4998	68001.331235	5.534388	7.130144	5.44	42625.620156	1.198657e+06	USS Wallace\nFPO AE 73316
4999	65510.581804	5.992305	6.792336	4.07	46501.283803	1.298950e+06	37778 George Ridges Apt. 509\nEast Holly, NV 2...

5000 rows × 7 columns

```
In [ ]:
```

Check info for any null values

- **we use the info() function in pandas to check all data values**

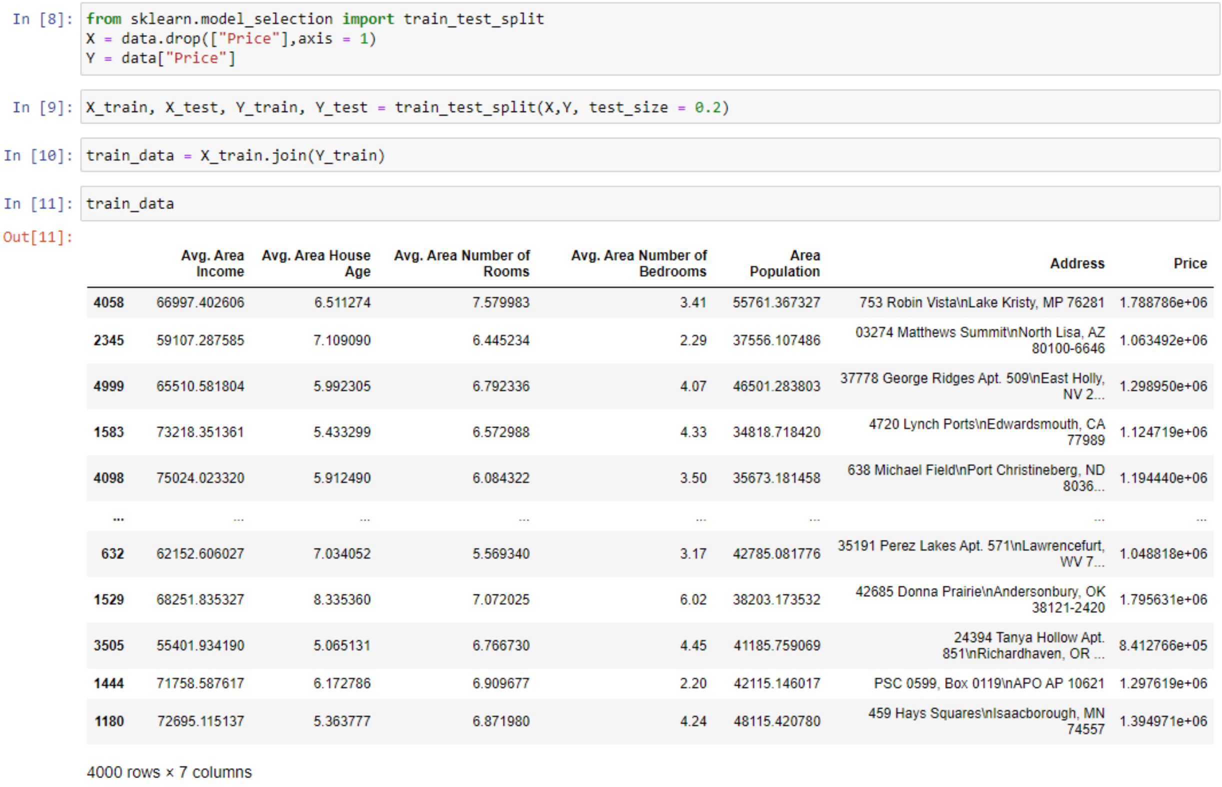


```
In [6]: data.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 5000 entries, 0 to 4999
        Data columns (total 7 columns):
         #   Column                        Non-Null Count  Dtype  
        ---  ------                        --------------  -----  
         0   Avg. Area Income              5000 non-null   float64
         1   Avg. Area House Age           5000 non-null   float64
         2   Avg. Area Number of Rooms     5000 non-null   float64
         3   Avg. Area Number of Bedrooms  5000 non-null   float64
         4   Area Population               5000 non-null   float64
         5   Price                         5000 non-null   float64
         6   Address                       5000 non-null   object 
        dtypes: float64(6), object(1)
        memory usage: 273.6+ KB

In [ ]:
```

- **Since the output showed no Null values we are free to proceed.**

<u>Splitting data into training and testing set</u>

- **we drop the target variable from the dataset and set it to X and set Y with the target variable data. (Target variable: Price)**
- **then we split both X and Y into training and testing sets.**
- **Finally we join our training set for X and Y then store it in train_data**



```
In [8]: from sklearn.model_selection import train_test_split
        X = data.drop(["Price"],axis = 1)
        Y = data["Price"]
```

```
In [9]: X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.2)
```

```
In [10]: train_data = X_train.join(Y_train)
```

```
In [11]: train_data
```

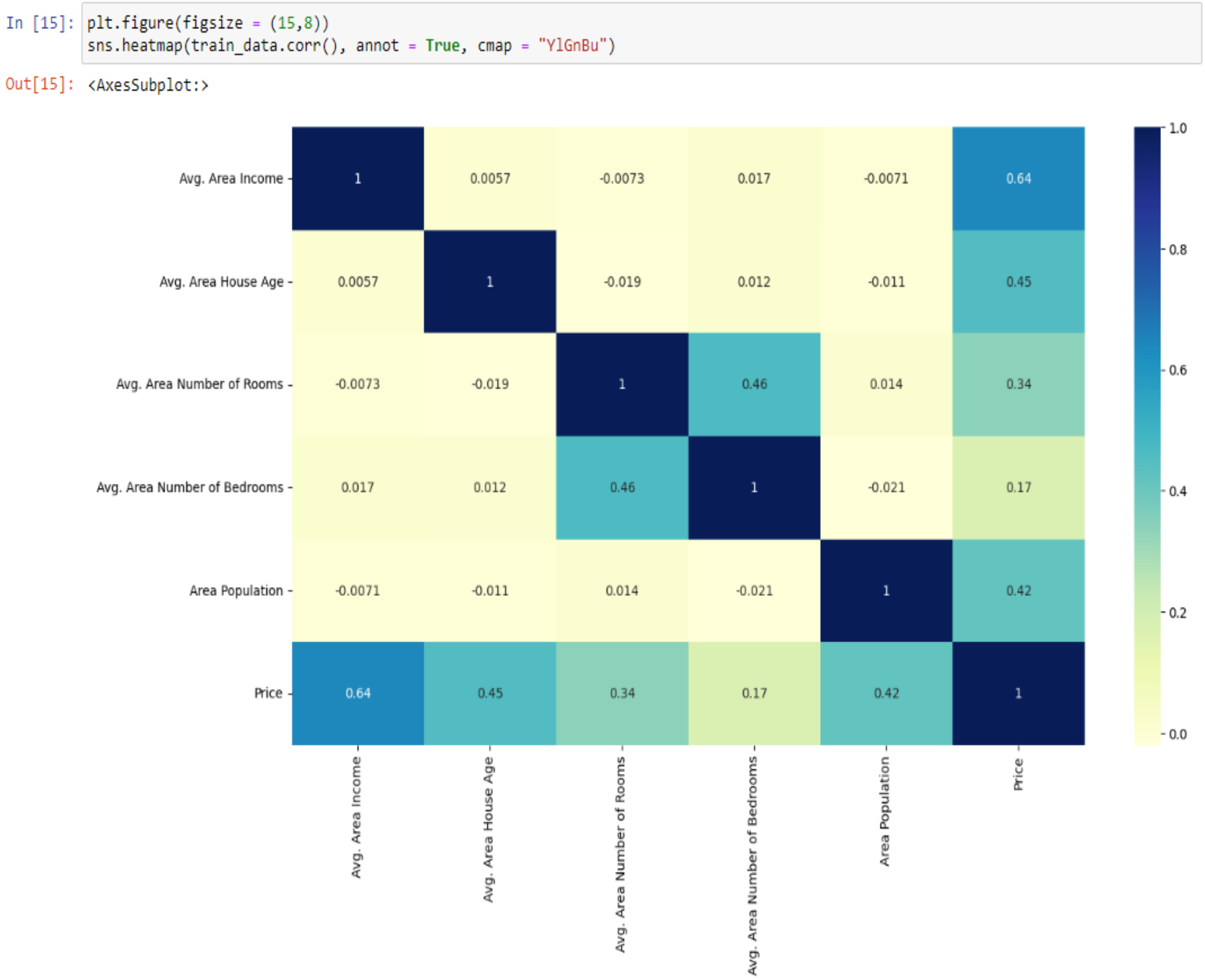
Out[11]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Address	Price
4058	66997.402606	6.511274	7.579983	3.41	55761.367327	753 Robin Vista\nLake Kristy, MP 76281	1.788786e+06
2345	59107.287585	7.109090	6.445234	2.29	37556.107486	03274 Matthews Summit\nNorth Lisa, AZ 80100-6646	1.063492e+06
4999	65510.581804	5.992305	6.792336	4.07	46501.283803	37778 George Ridges Apt. 509\nEast Holly, NV 2...	1.298950e+06
1583	73218.351361	5.433299	6.572988	4.33	34818.718420	4720 Lynch Ports\nEdwardsmouth, CA 77989	1.124719e+06
4098	75024.023320	5.912490	6.084322	3.50	35673.181458	638 Michael Field\nPort Christineberg, ND 8036...	1.194440e+06
...	...	...	...	...	...	...	...
632	62152.606027	7.034052	5.569340	3.17	42785.081776	35191 Perez Lakes Apt. 571\nLawrencefurt, WV 7...	1.048818e+06
1529	68251.835327	8.335360	7.072025	6.02	38203.173532	42685 Donna Prairie\nAndersonbury, OK 38121-2420	1.795631e+06
3505	55401.934190	5.065131	6.766730	4.45	41185.759069	24394 Tanya Hollow Apt. 851\nRichardhaven, OR ...	8.412766e+05
1444	71758.587617	6.172786	6.909677	2.20	42115.146017	PSC 0599, Box 0119\nAPO AP 10621	1.297619e+06
1180	72695.115137	5.363777	6.871980	4.24	48115.420780	459 Hays Squares\nIsaacborough, MN 74557	1.394971e+06

4000 rows × 7 columns

Finding Correlation between all data with target variable (“Price”)

- **We use the heatmap() function from seaborn to visualize the correlation between the data and the target variable “Price”, we pass in the correlation matrix of train_data as the parameter.**



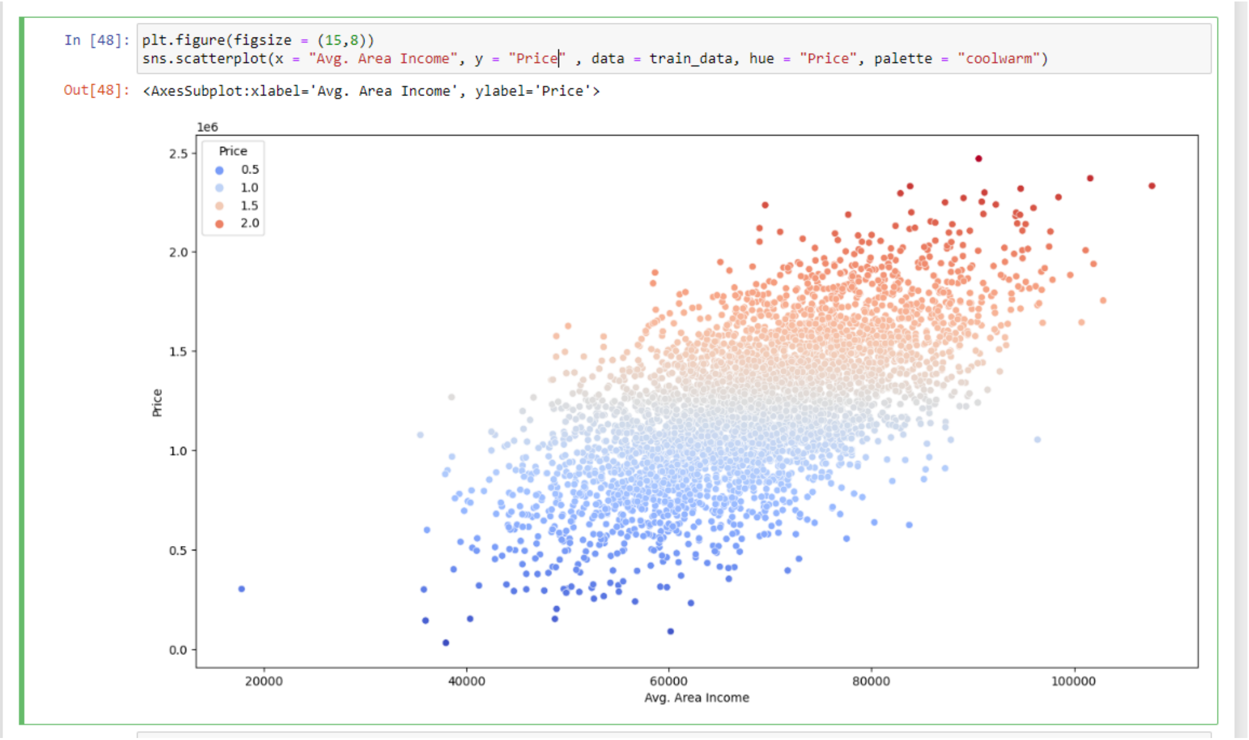
```
In [15]: plt.figure(figsize = (15,8))
         sns.heatmap(train_data.corr(), annot = True, cmap = "YlGnBu")

Out[15]: <AxesSubplot:>
```

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
Avg. Area Income	1	0.0057	-0.0073	0.017	-0.0071	0.64
Avg. Area House Age	0.0057	1	-0.019	0.012	-0.011	0.45
Avg. Area Number of Rooms	-0.0073	-0.019	1	0.46	0.014	0.34
Avg. Area Number of Bedrooms	0.017	0.012	0.46	1	-0.021	0.17
Area Population	-0.0071	-0.011	0.014	-0.021	1	0.42
Price	0.64	0.45	0.34	0.17	0.42	1

- **we can see that Area population correlates highly with the target variable “Price”**

Plotting the data to check correlation

- Using the scatterplot() function in seaborn we plot the data between Avg. Area Income and Price to find out its correlation.



```
In [48]: plt.figure(figsize = (15,8))
         sns.scatterplot(x = "Avg. Area Income", y = "Price" , data = train_data, hue = "Price", palette = "coolwarm")

Out[48]: <AxesSubplot:xlabel='Avg. Area Income', ylabel='Price'>
```

- from the scatter plot we can see that the Avg. Area Income plays a huge role in the price value of a house.

Conclusion

Thus, the dataset was cleaned and preprocessed and the target variable was assessed. We split the dataset into training and testing sets and upon analysis, we found the high correlation between Avg. Area Income and the target variable "Price". We also found variables that did not play much of a role in assessing the target variable "Price"