Московский государственный технический университет им. Н.Э. Баумана.

Факультет «Информатика и управление»

Кафедра «Системы обработки информации и управления»

Курс «Теория машинного обучения»

Отчет по лабораторной работе №1

Выполнил:
студент группы ИУ5-64
Кузьмин Роман

Подпись и дата:

Проверил:
преподаватель каф. ИУ5
Гапанюк Ю.Е.

Подпись и дата:

Москва, 2022 г.

Описание задания

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из Scikit-learn.
- Пример преобразования датасетов Scikit-learn в Pandas Dataframe можно посмотреть здесь.
- Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

Датасет

_california_housing_dataset:

California Housing dataset

Data Set Characteristics:

:Number of Instances: 20640

:Number of Attributes: 8 numeric, predictive attributes and the target

:Attribute Information:
- MedInc median income in block group
- HouseAge median house age in block group
- AveRooms average number of rooms per household
- AveBedrms average number of bedrooms per household
- Population block group population
- AveOccup average number of household members
- Latitude block group latitude
- Longitude block group longitude

:Missing Attribute Values: None

This dataset was obtained from the StatLib repository.
https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html

The target variable is the median house value for California districts,
expressed in hundreds of thousands of dollars ($100,000).

This dataset was derived from the 1990 U.S. census, using one row per census block group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people).

An household is a group of people residing within a home. Since the average number of rooms and bedrooms in this dataset are provided per household, these
columns may take surpinsingly large values for block groups with few households and many empty houses, such as vacation resorts.

Текст программы

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
 from sklearn.datasets import fetch_california_housing
data = fetch_california_housing(as_frame=True)['data']
data["price"] = fetch_california_housing(as_frame=True)['target']
 data.head()
 data.dtypes
 # Проверим наличие пустых значений
# Цикл по колонкам датасета
for col in data.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))
 data.describe()
 # Определим уникальные значения для целевого признака
data['price'].unique()
 fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='AveRooms', y='AveBedrms', data=data)
 fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['price'])
 sns.jointplot(x='AveRooms', y='AveBedrms', data=data)
 sns.pairplot(data)
 sns.boxplot(x=data['price'])
 sns.violinplot(x=data['HouseAge'])
 sns.heatmap(data.corr(), cmap='magma', annot=True, fmt='.2f')
```

Анализ результатов

```
data.head()
```

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	price
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

```
data.dtypes
```

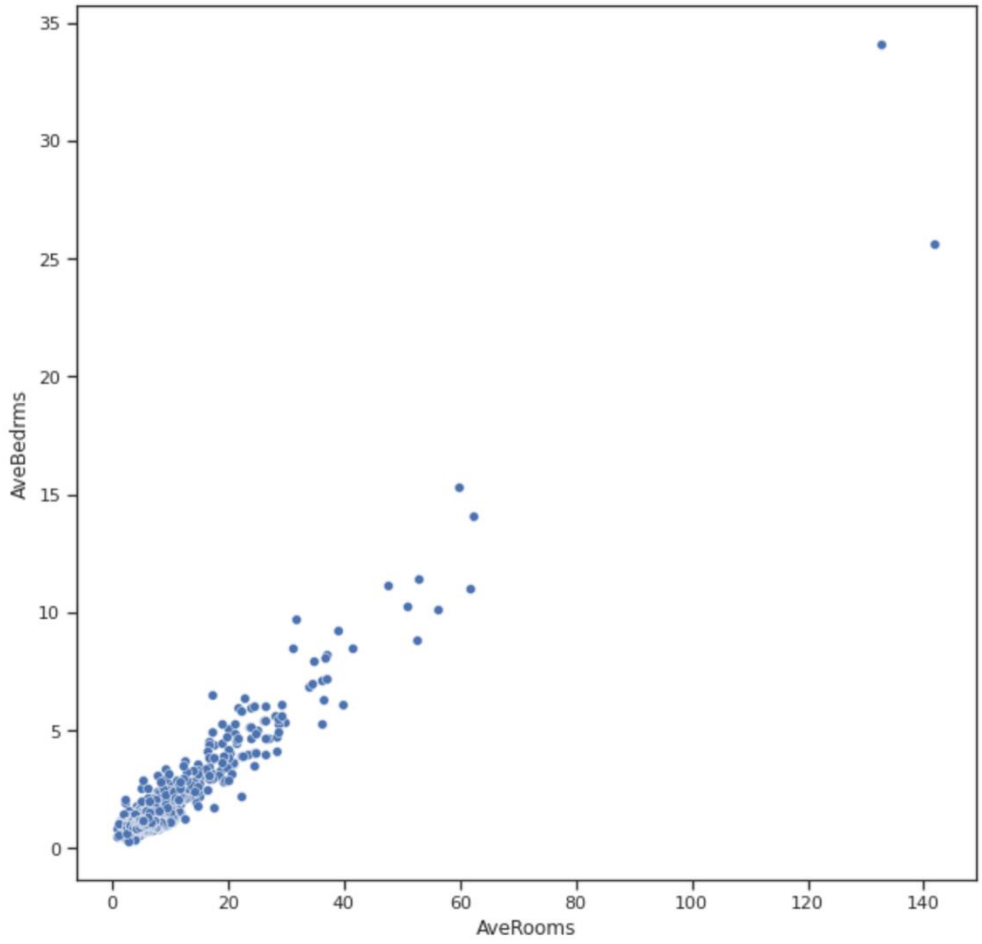
```
MedInc        float64
HouseAge      float64
AveRooms      float64
AveBedrms     float64
Population    float64
AveOccup      float64
Latitude      float64
Longitude     float64
price         float64
dtype: object
```

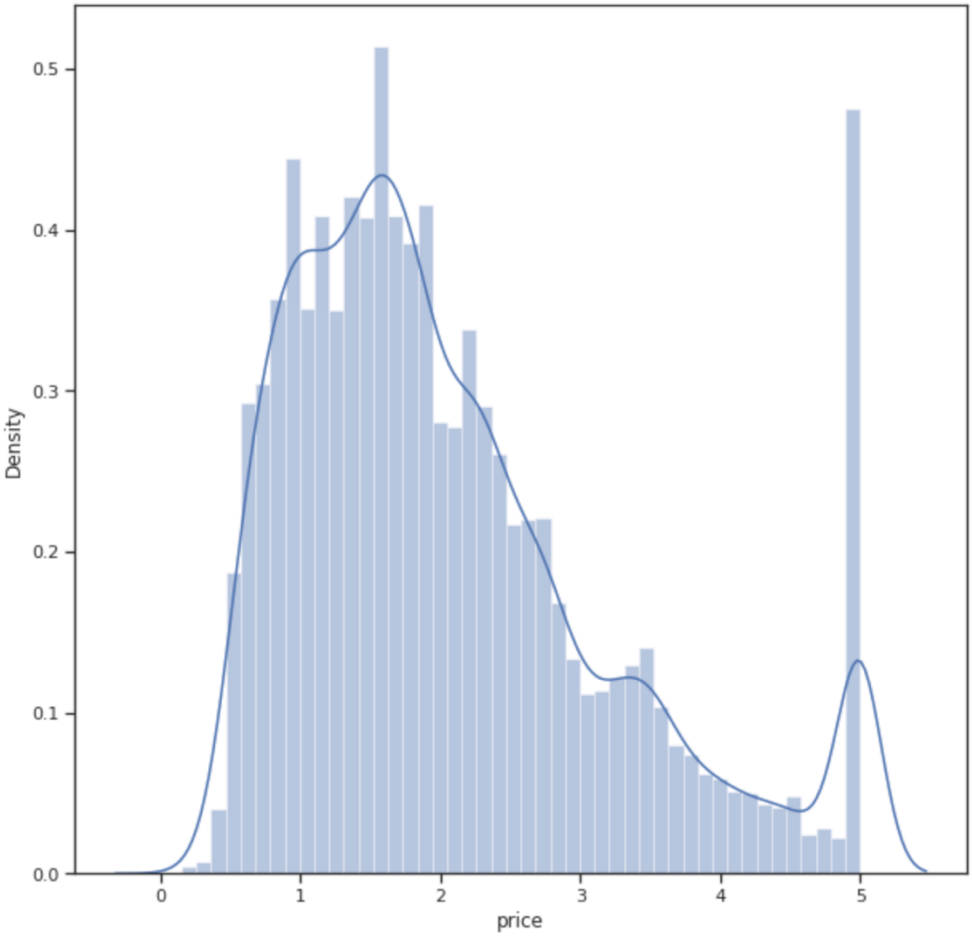
```
MedInc - 0
HouseAge - 0
AveRooms - 0
AveBedrms - 0
Population - 0
AveOccup - 0
Latitude - 0
Longitude - 0
price - 0
```

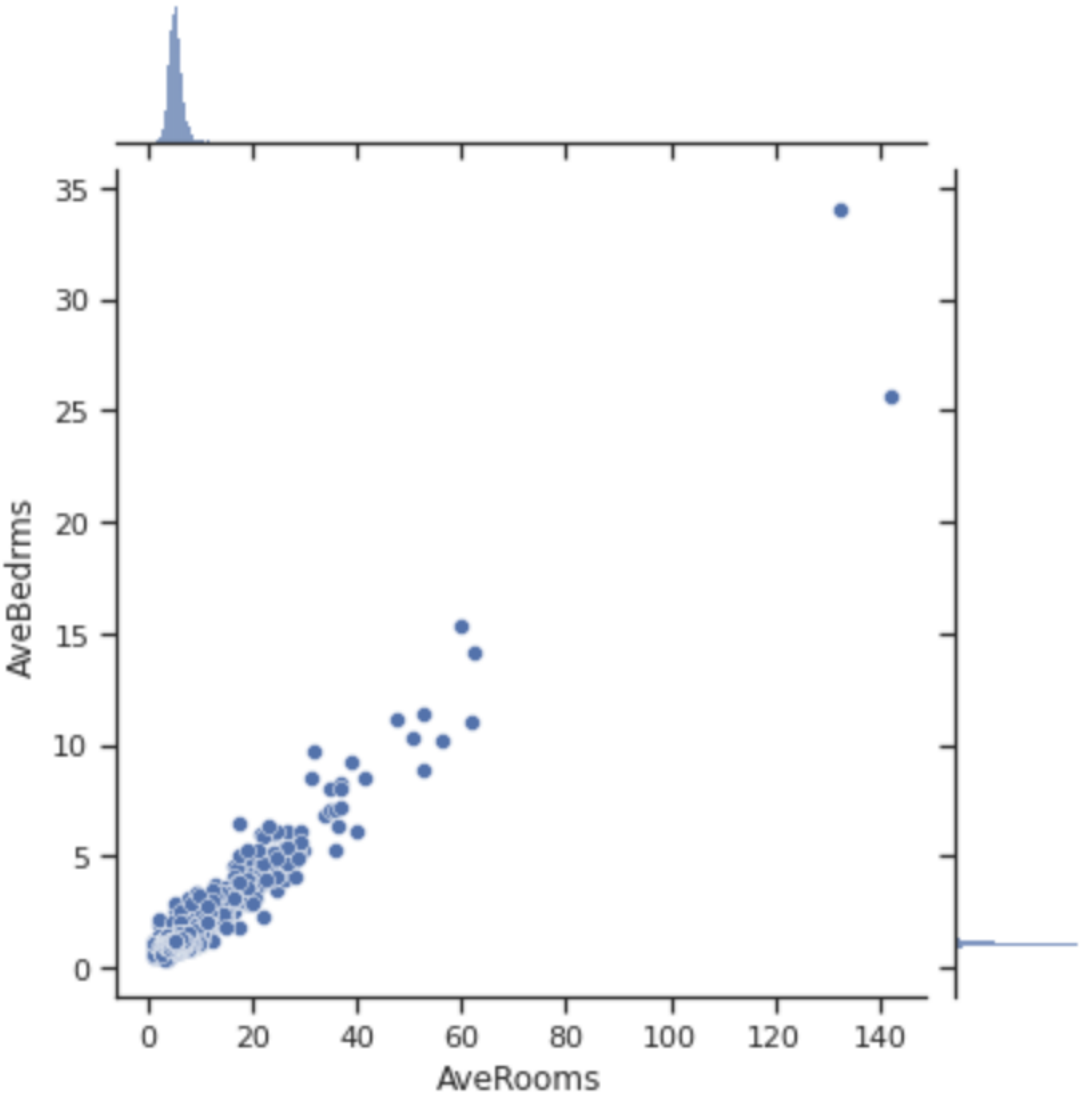
```
data.describe()
```

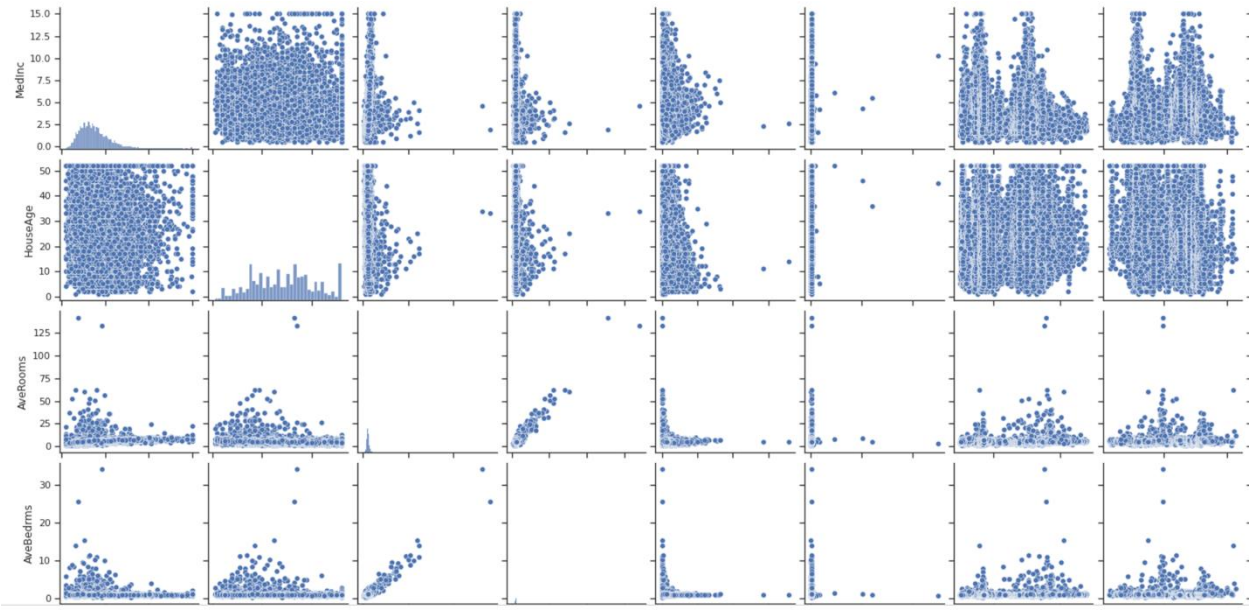
	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	price
count	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000
mean	3.870671	28.639486	5.429000	1.096675	1425.476744	3.070655	35.631861	-119.569704	2.068558
std	1.899822	12.585558	2.474173	0.473911	1132.462122	10.386050	2.135952	2.003532	1.153956
min	0.499900	1.000000	0.846154	0.333333	3.000000	0.692308	32.540000	-124.350000	0.149990
25%	2.563400	18.000000	4.440716	1.006079	787.000000	2.429741	33.930000	-121.800000	1.196000
50%	3.534800	29.000000	5.229129	1.048780	1166.000000	2.818116	34.260000	-118.490000	1.797000
75%	4.743250	37.000000	6.052381	1.099526	1725.000000	3.282261	37.710000	-118.010000	2.647250
max	15.000100	52.000000	141.909091	34.066667	35682.000000	1243.333333	41.950000	-114.310000	5.000010

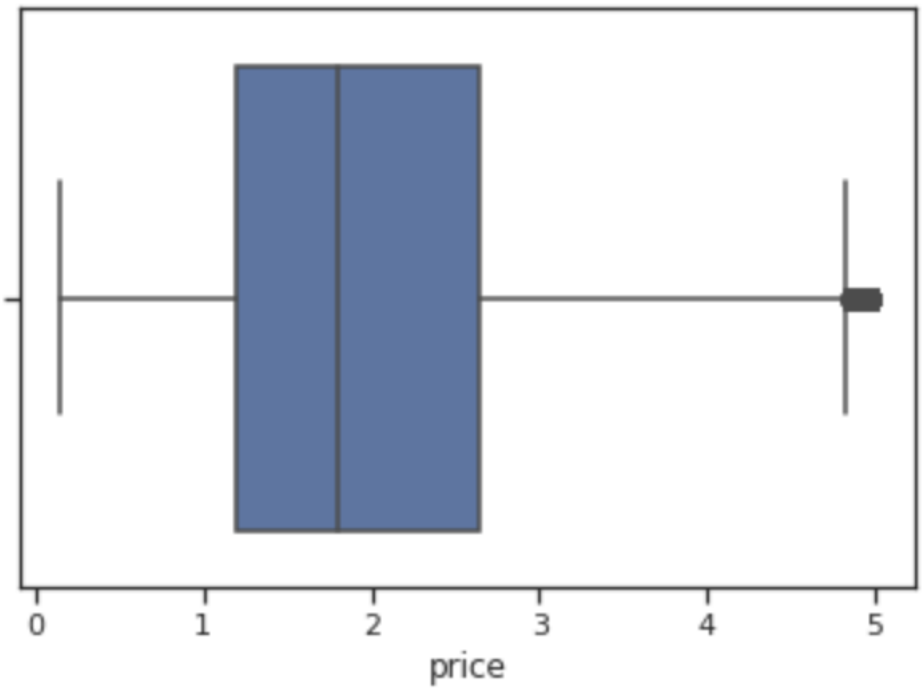
```
array([4.526, 3.585, 3.521, ..., 4.258, 2.007, 0.47 ])
```


AveBedrms
AveRooms
0
5
10
15
20
25
30
35
0
20
40
60
80
100
120
140


Density
price
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5

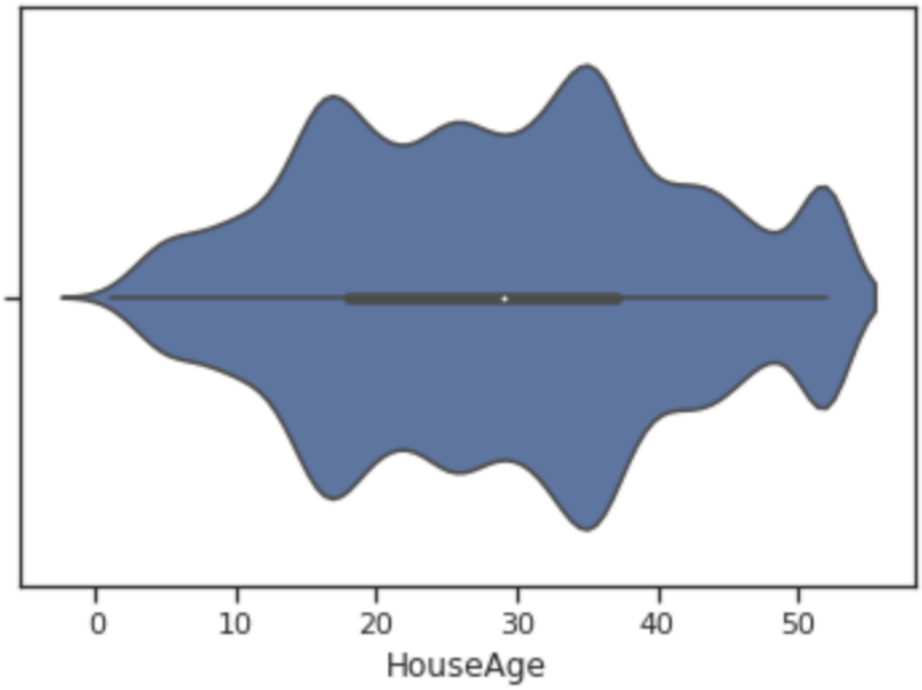

AveBedrms
AveRooms
0
5
10
15
20
25
30
35
0
20
40
60
80
100
120
140

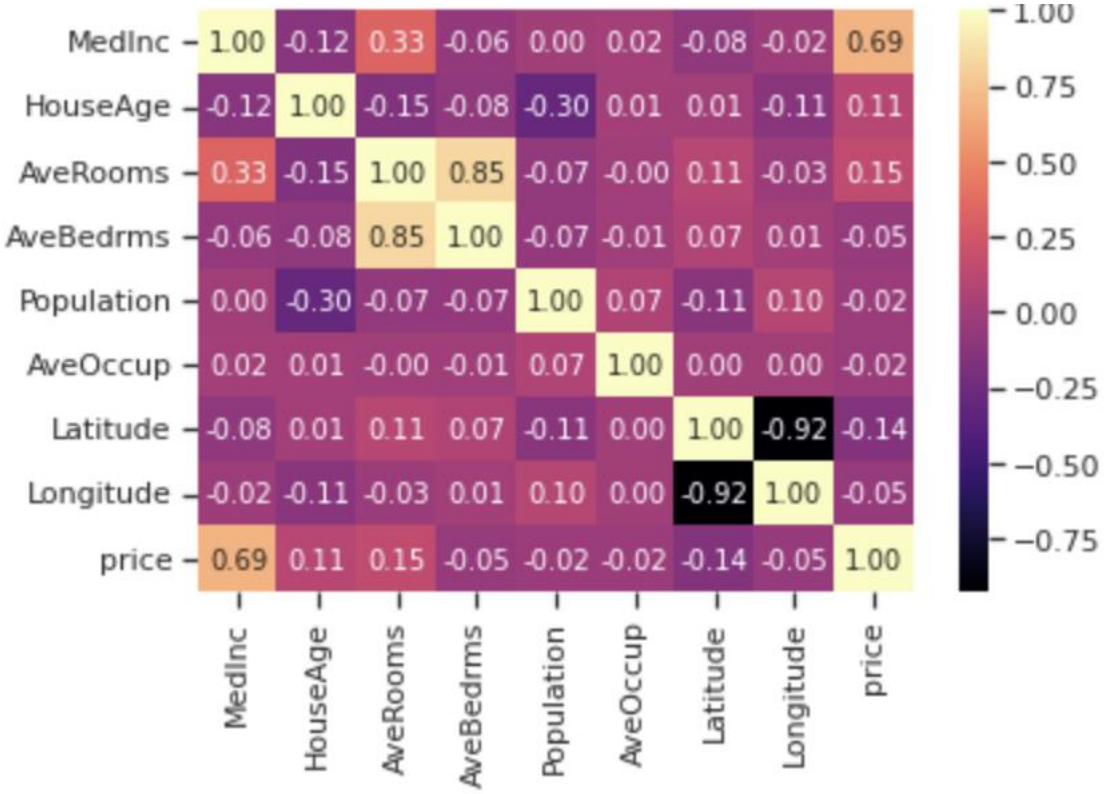

MedInc
HouseAge
AveRooms
AveBedrms




```
sns.violinplot(x=data['HouseAge'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0
```




MedInc 1.00 -0.12 0.33 -0.06 0.00 0.02 -0.08 -0.02 0.69
HouseAge -0.12 1.00 -0.15 -0.08 -0.30 0.01 0.01 -0.11 0.11
AveRooms 0.33 -0.15 1.00 0.85 -0.07 -0.00 0.11 -0.03 0.15
AveBedrms -0.06 -0.08 0.85 1.00 -0.07 -0.01 0.07 0.01 -0.05
Population 0.00 -0.30 -0.07 -0.07 1.00 0.07 -0.11 0.10 -0.02
AveOccup 0.02 0.01 -0.00 -0.01 0.07 1.00 0.00 0.00 -0.02
Latitude -0.08 0.01 0.11 0.07 -0.11 0.00 1.00 -0.92 -0.14
Longitude -0.02 -0.11 -0.03 0.01 0.10 0.00 -0.92 1.00 -0.05
price 0.69 0.11 0.15 -0.05 -0.02 -0.02 -0.14 -0.05 1.00
MedInc
HouseAge
AveRooms
AveBedrms
Population
AveOccup
Latitude
Longitude
price
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75