Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления (ИУ)

КАФЕДРА Система обработки информации и управления

ДИСЦИПЛИНА Методы машинного обучения

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 1

Создание "истории о данных" (Data Storytelling)

название лабораторной работы

Группа ИУ5-14М

Студент 14.03.2022 _____ Молева А. А.
дата выполнения работы *подпись* *фамилия, и.о.*

Преподаватель _____ Гапанюк Ю. Е.
подпись *фамилия, и.о.*

Москва, 2022 г.

Цель работы

Цель лабораторной работы: изучение различных методов визуализация данных и создание истории на основе данных.

Краткое описание. Построение графиков, помогающих понять структуру данных, и их интерпретация.

Задание

- Выбрать набор данных (датасет).
- Создать "историю о данных" в виде юпитер-ноутбука, с учетом следующих требований:
  - История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
  - На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
  - Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
  - Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
  - История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.
- Сформировать отчет и разместить его в своем репозитории на github.

Текст программы

```
dataset = load_boston()
import pandas as pd
X = pd.DataFrame(dataset['data'], columns=dataset['feature_names'])
y = pd.Series(dataset['target'], name='target')
from sklearn.datasets import fetch_california_housing
dataset = fetch_california_housing(return_X_y=True, as_frame=True)
X, y = dataset
X.info()
X.describe()
import pandas as pd
housing = pd.concat([X, y], axis=1)
from matplotlib import pyplot as plt

#skatter
housing.plot(kind='scatter', x='Longitude', y='Latitude', alpha=0.4, s=X.Popu
lation/100, label='population', figsize=(10,7), c=housing.MedHouseVal,
             cmap=plt.get_cmap("jet"), colorbar=True)
plt.legend()


#парные диаграммы
import seaborn as sns
sns.pairplot(housing)

# Violin plot

fig, ax = plt.subplots(9, 1, figsize=(10, 35))
for idx, column in enumerate(housing.columns):
    sns.violinplot(x=housing[column], ax=ax[idx])

# Histograms
fig, ax = plt.subplots(9, 1, figsize=(10, 35))
for idx, column in enumerate(housing.columns):
    sns.distplot(x=housing[column], ax=ax[idx], axlabel=column)

# Boxplot
fig, ax = plt.subplots(9, 1, figsize=(10, 35))
for idx, column in enumerate(housing.columns):
    sns.boxplot(y=housing[column], ax=ax[idx])

housing.corr()
# Вывод значений в ячейках
plt.figure(figsize=(10,10))
sns.heatmap(housing.corr(), annot=True, fmt='.3f')
```

Экранные формы

X

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33
...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1.0	273.0	21.0	391.99	9.67
502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1.0	273.0	21.0	396.90	9.08
503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1.0	273.0	21.0	396.90	5.64
504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1.0	273.0	21.0	393.45	6.48
505	0.04741	0.0	11.93	0.0	0.573	6.030	80.8	2.5050	1.0	273.0	21.0	396.90	7.88

506 rows × 13 columns

Рисунок 1 – Датасет

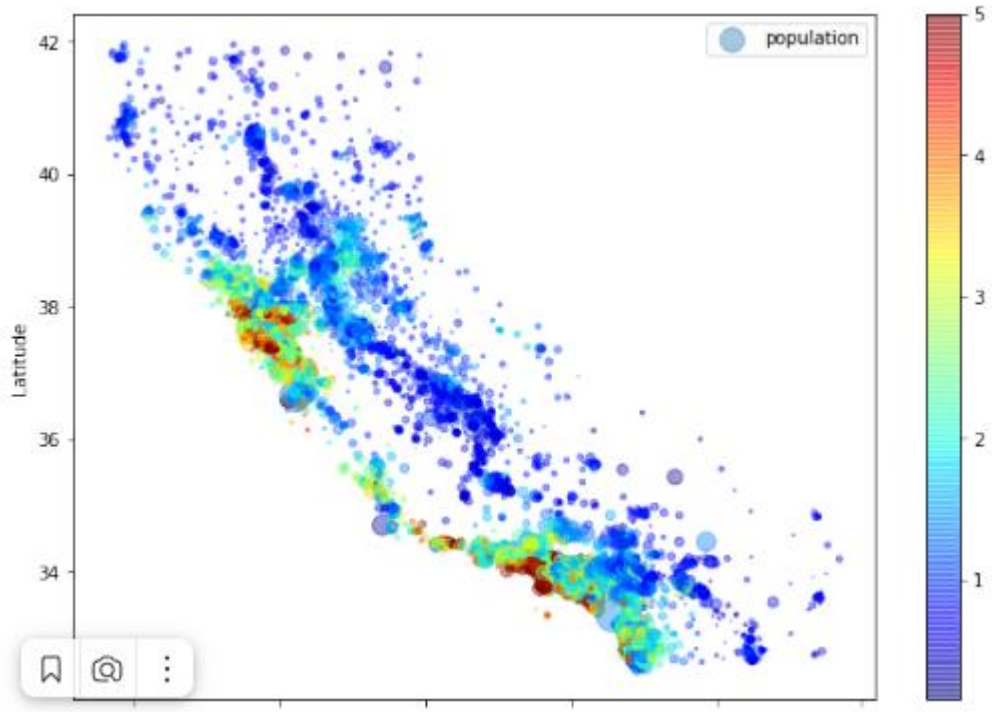

Рисунок 2 – Диаграмма рассеяния

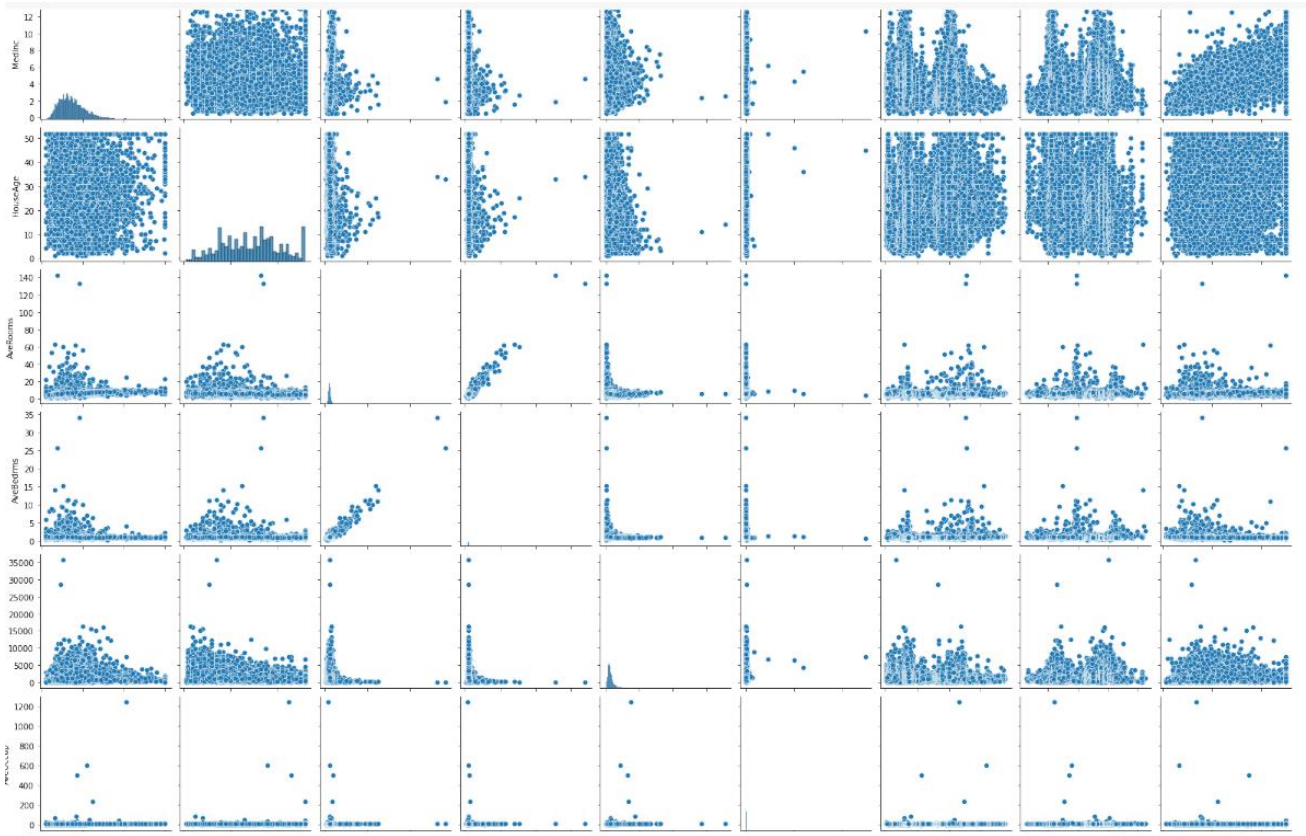

Рисунок 3 – Парные диаграммы

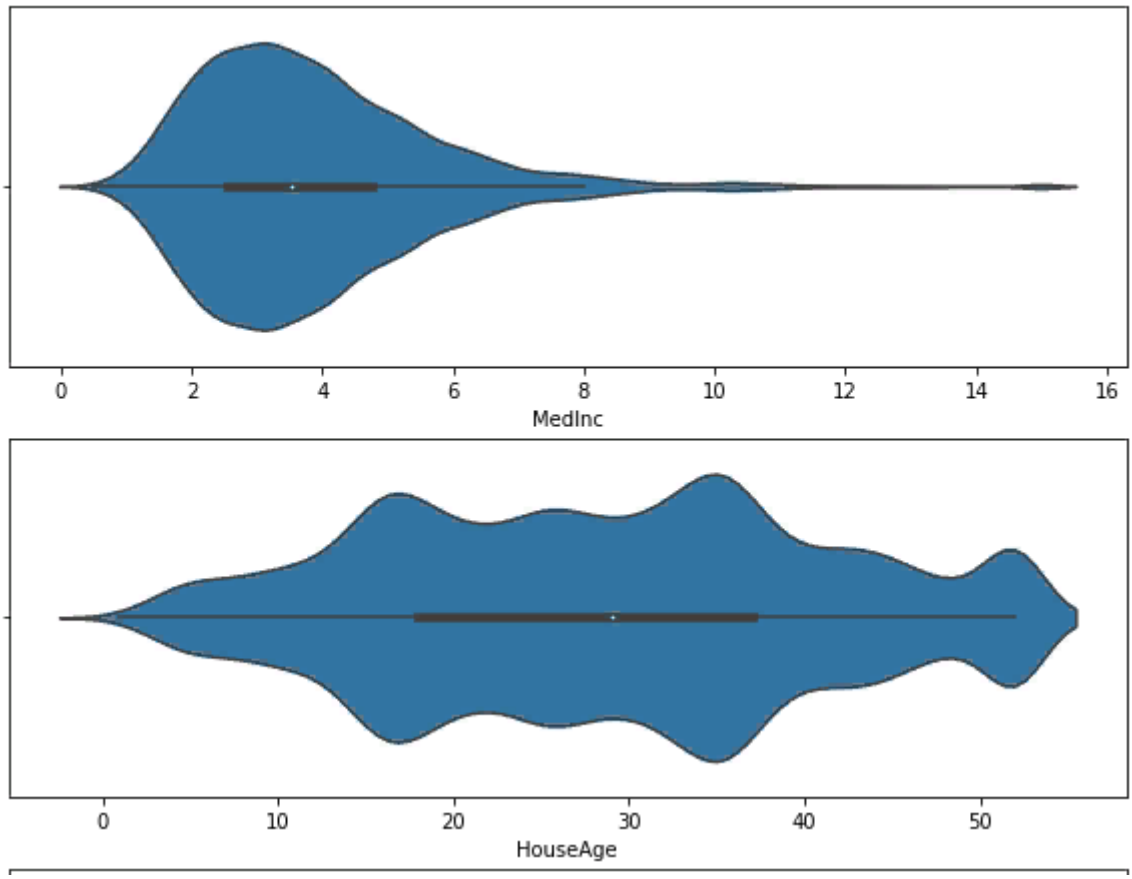

Рисунок 4 – Скрипичная диаграмма

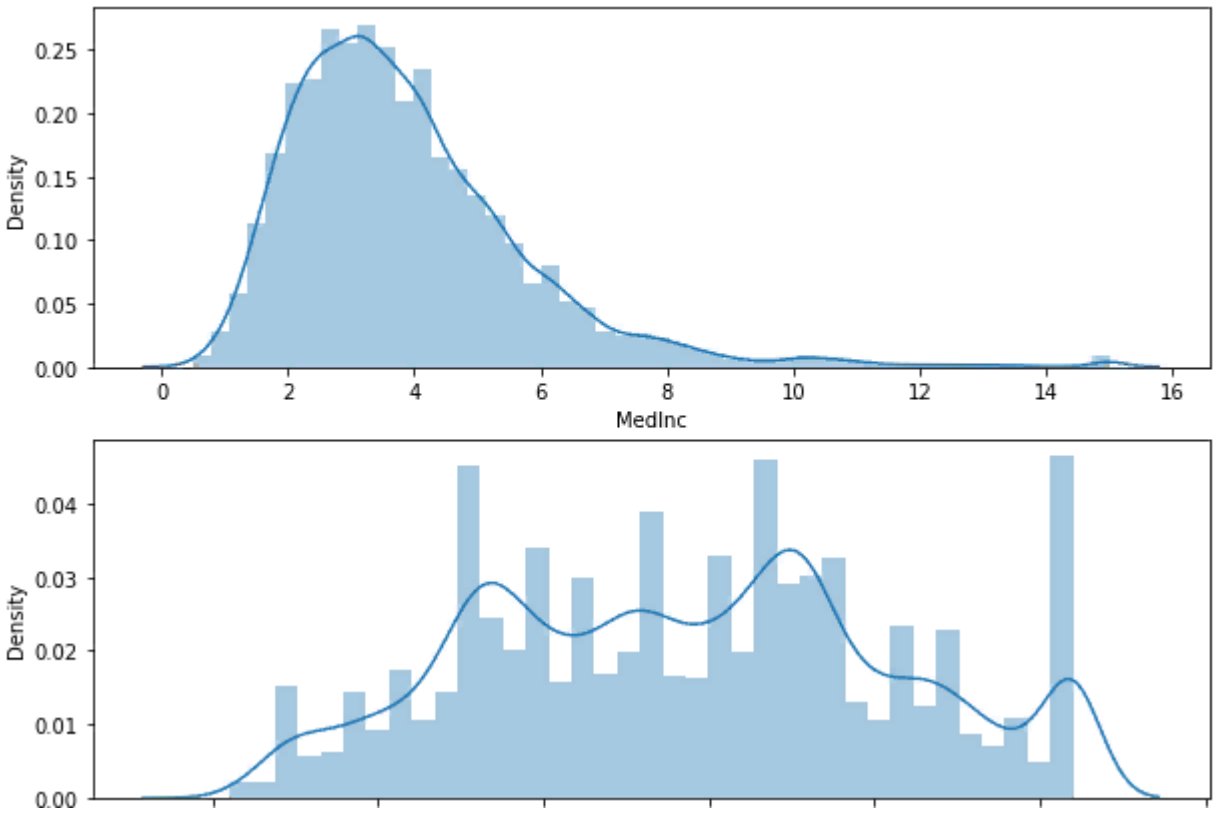

Рисунок 5 – Гистограммы

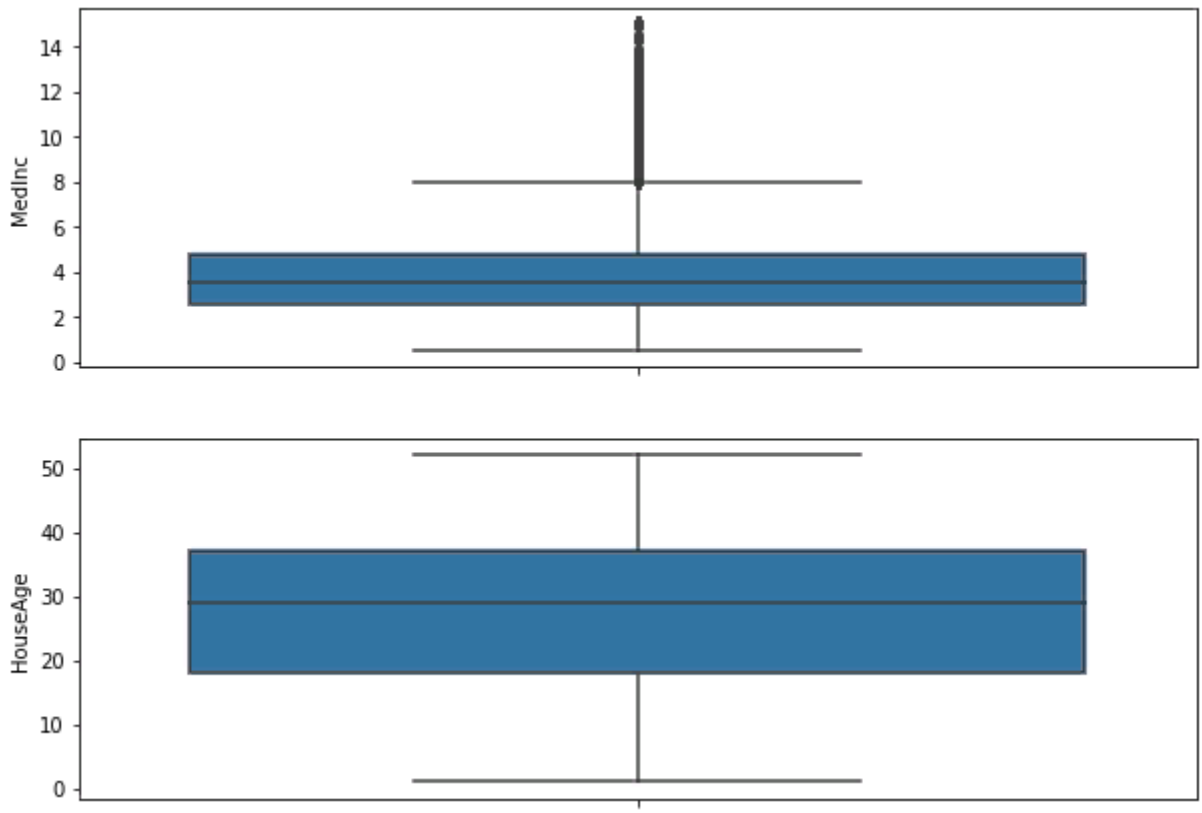

Рисунок 6 – Боксплот

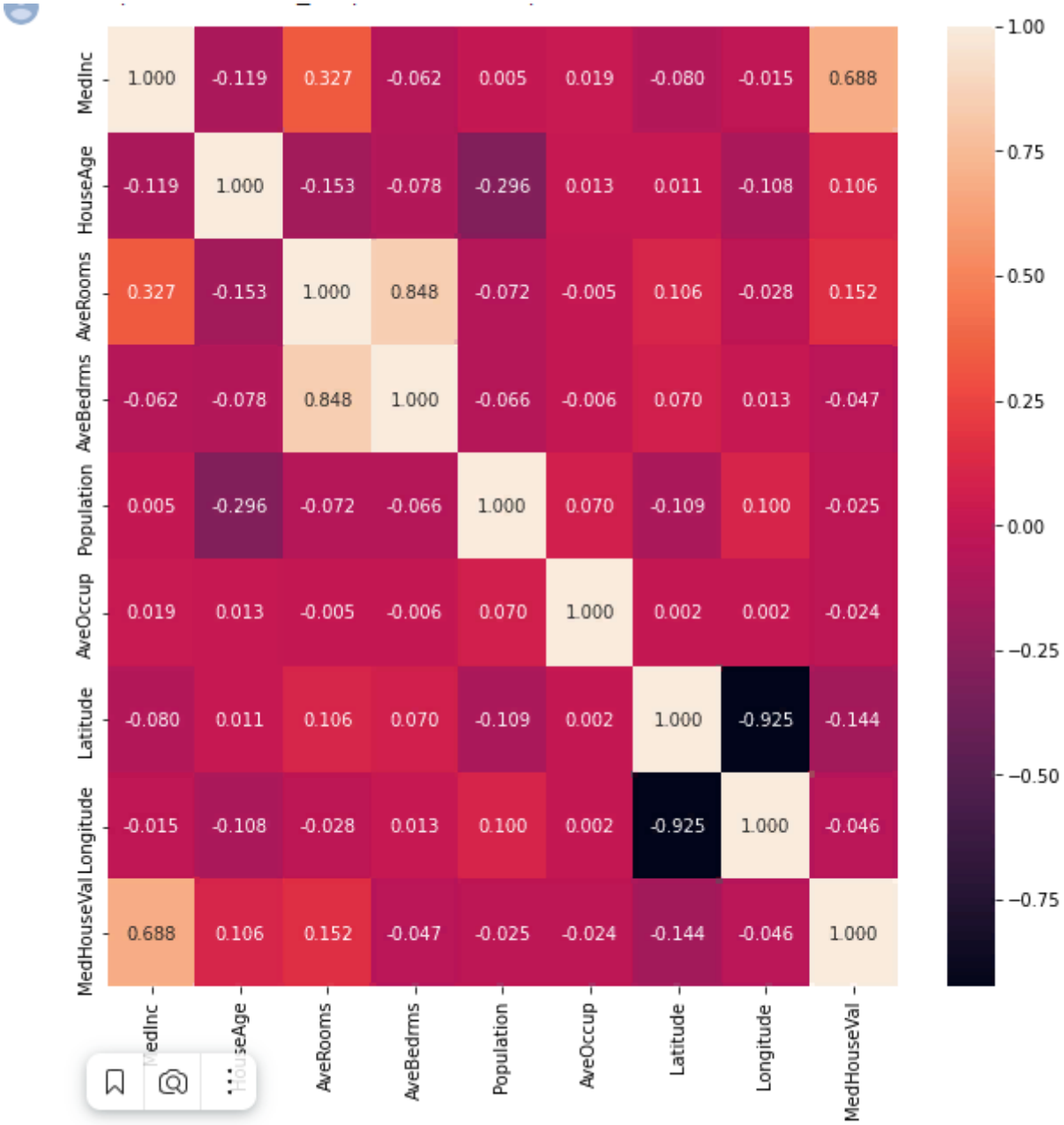

Рисунок 7 – Тепловая карта

Выводы

В результате проделанной работы была создана история о данных со следующими диаграммами: диаграмма расеяния, парные диаграммы, скрипичные диаграммы, гистограммы, боксплоты, тепловая карта.