Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Системы обработки информации и управления»

Лабораторная работа №1
по курсу «Методы машинного обучения»

«Создание "истории о данных" (Data Storytelling).»

ИСПОЛНИТЕЛЬ:

Ерохин И.А.
Группа ИУ5-24М

_____

"__"_____2022 г.

Москва 2022

Цель работы:

Изучение различных методов визуализация данных и создание истории на основе данных.

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.

Выполнение:

Был выбран датасет о качестве вина:

In [3]:
```
data = pd.read_csv('WineQT.csv', sep=",")
```

In [4]:
```
data.head()
```

Out[4]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	1
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	2
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	3
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	4

In [5]:
```
data.shape
```

Out[5]:
```
(1143, 13)
```

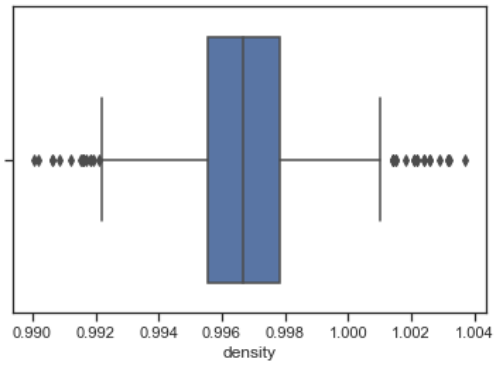
In [6]:
```
data.dtypes
```

Out[6]:
```
fixed acidity           float64
volatile acidity        float64
citric acid             float64
residual sugar          float64
chlorides               float64
free sulfur dioxide     float64
total sulfur dioxide    float64
density                 float64
pH                      float64
sulphates               float64
alcohol                 float64
quality                   int64
Id                        int64
dtype: object
```

Были построены графики «ящик с усами» для колонок density, fixed acidity

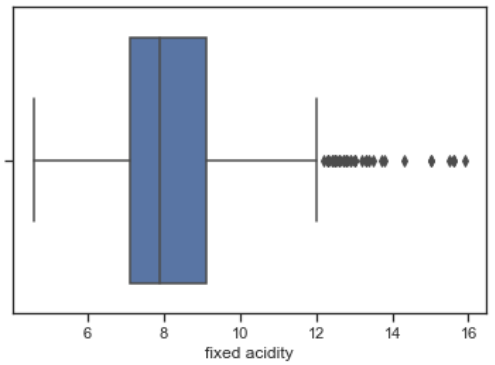
In [17]:
```
sns.boxplot(x=data['density'])
```

Out[17]:
```
<AxesSubplot:xlabel='density'>
```



In [21]:
```
sns.boxplot(x=data['fixed acidity'])
```

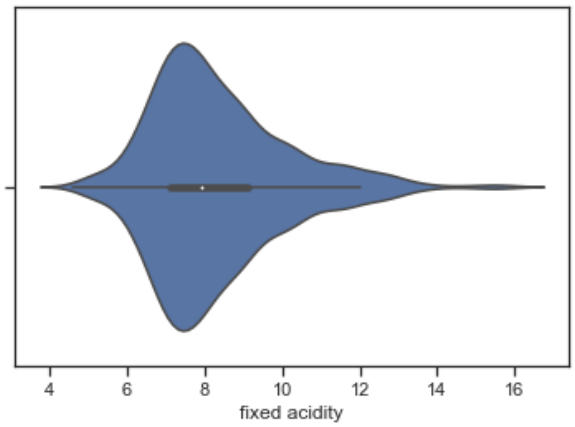
Out[21]:
```
<AxesSubplot:xlabel='fixed acidity'>
```



Для уточнения данных колонки fixed acidity было принято решение добавить на диаграмму точки, но из-за размеров датасета данные оказались трудны для чтения и пришлось перейти к построению скрипичного графика

```
In [27]: sns.violinplot(x=data['fixed acidity'])

Out[27]: <AxesSubplot:xlabel='fixed acidity'>
```

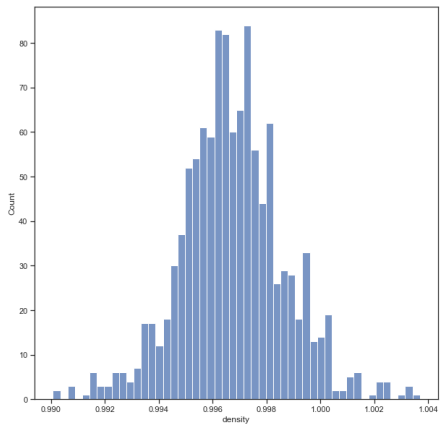



```
In [ ]:
```

Далее было проведено изучение гистограмм для аналогичных колонок

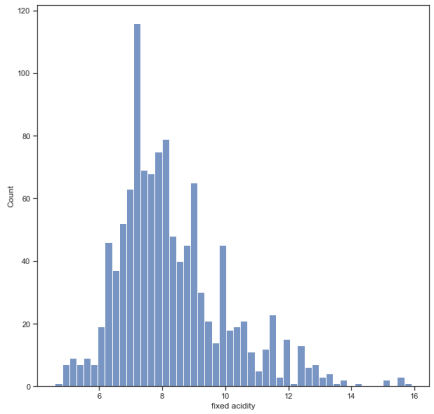
```
In [13]: fig, ax = plt.subplots(figsize=(10,10))
         sns.histplot(data['density'], bins = 50)

Out[13]: <AxesSubplot:xlabel='density', ylabel='Count'>
```




```
fig, ax = plt.subplots(figsize=(10,10))
sns.histplot(data['fixed acidity'], bins = 50)

ut[11]: <AxesSubplot:xlabel='fixed acidity', ylabel='Count'>
```

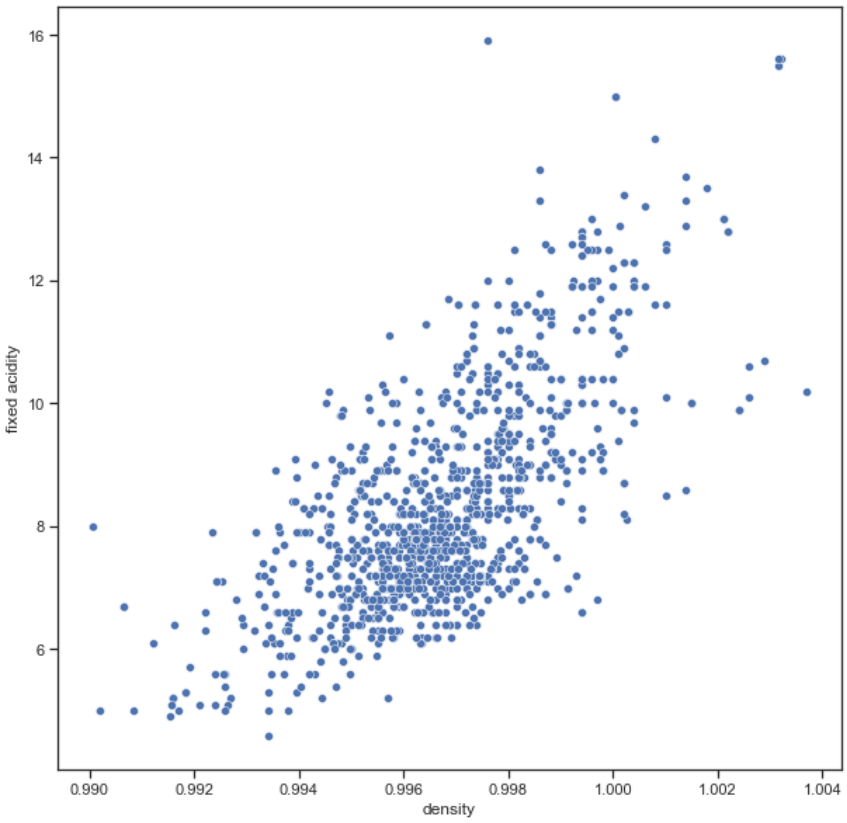


Для density можем наблюдать колоколоподобный график, а, следовательно, нормальное распределение. Колонка fixed acid требует нормализации.

Далее была рассмотрена диаграмма рассеяния. Наблюдается слабая прямая корреляция.

In [12]:
```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='density', y='fixed acidity',
                data=data)
```

Out[12]: <AxesSubplot:xlabel='density', ylabel='fixed acidity'>



In [ ]:
```
# можем наблюдать что присутствует прямая корреляция
```

Вывод:

В результате работы был выбран датасет и сформирована история данных в виде юпитер-ноутбука, соответствующая требованиями задания. Были изучены различные методы визуализация данных.