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

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему
«Создание истории о данных»

Выполнил:
студент группы ИУ5-24М
Подопригорова С. С.

_____

Москва — 2023 г.

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:
    1. История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
    2. На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных “неудачных” графиков.
    3. Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
    4. 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.
    5. 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.

1. Описание набора данных

Мы будем использовать набор данных по распознаванию вин https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

Эти данные являются результатами химического анализа вин, выращенных в одном регионе Италии тремя разными культиваторами. Было проведено тринадцать различных измерений для различных компонентов, содержащихся в трех типах вина.

Каждый файл содержит следующие колонки: - fixed acidity - volatile acidity - citric acid - residual sugar - chlorides - free sulfur dioxide - total sulfur dioxide - density - pH - sulphates - alcohol - quality - целевой признак

[1]:
```
import numpy as np
import pandas as pd

import seaborn as sns
sns.set(style="ticks")

import matplotlib.pyplot as plt
%matplotlib inline
```

[2]:
```
data = pd.read_csv("wine.csv")
```

[31]:
```
data.head()
```

[31]:
```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.4              0.70         0.00             1.9      0.076
```

```
1            7.8              0.88         0.00             2.6      0.098
2            7.8              0.76         0.04             2.3      0.092
3           11.2              0.28         0.56             1.9      0.075
4            7.4              0.70         0.00             1.9      0.076

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 11.0                  34.0   0.9978  3.51       0.56
1                 25.0                  67.0   0.9968  3.20       0.68
2                 15.0                  54.0   0.9970  3.26       0.65
3                 17.0                  60.0   0.9980  3.16       0.58
4                 11.0                  34.0   0.9978  3.51       0.56

   alcohol  quality
0      9.4        0
1      9.8        0
2      9.8        0
3      9.8        1
4      9.4        0
```

[17]:
```
data.shape
```

[17]:
```
(1599, 12)
```

1599 строк, 7 колонок

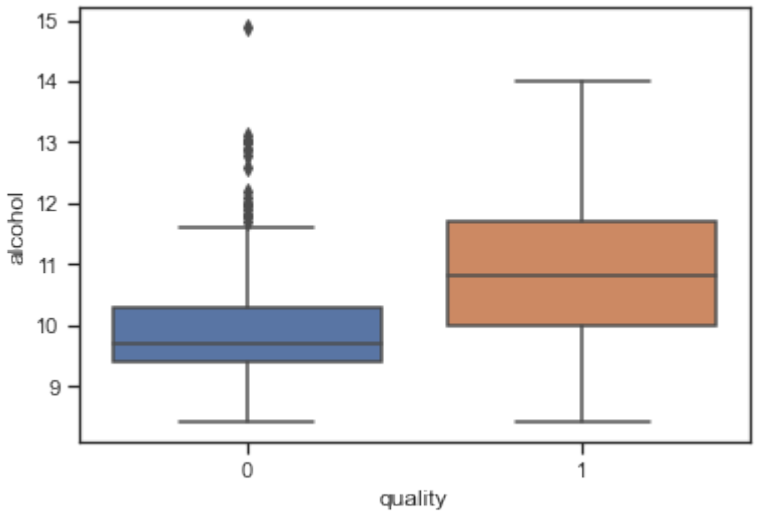
[3]:
```
data['quality'] = data['quality'].apply(lambda x: 1 if x == 'good' else 0)
```

2. Визуальное исследование датасета

Рассмотрим процент **алкоголя** в вине хорошего и плохого качества.

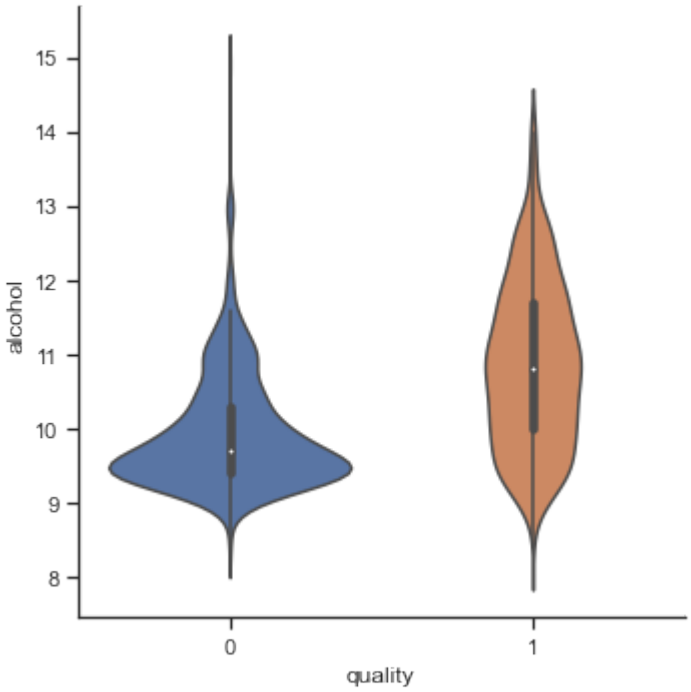
[5]:
```
sns.boxplot(x='quality', y='alcohol', data=data)
```

[5]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fab8c64fe80>
```



[4]:
```
sns.catplot(x="quality", y="alcohol", data=data, kind="violin")
```

[4]:
```
<seaborn.axisgrid.FacetGrid at 0x7fe8b025bd60>
```



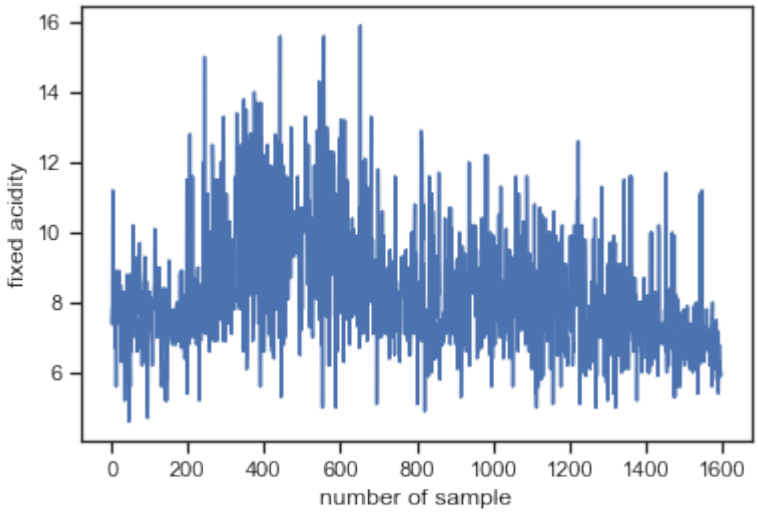
Видим различия в распределении процентности алкоголя среди качественных и некачественных вин. У качественных вин в среднем содержание алкоголя выше, при этом больше разброс.

Также видим выбросы в категории некачественных вин с завышенным содержанием алкоголя.

Рассмотрим **фиксированную кислотность**

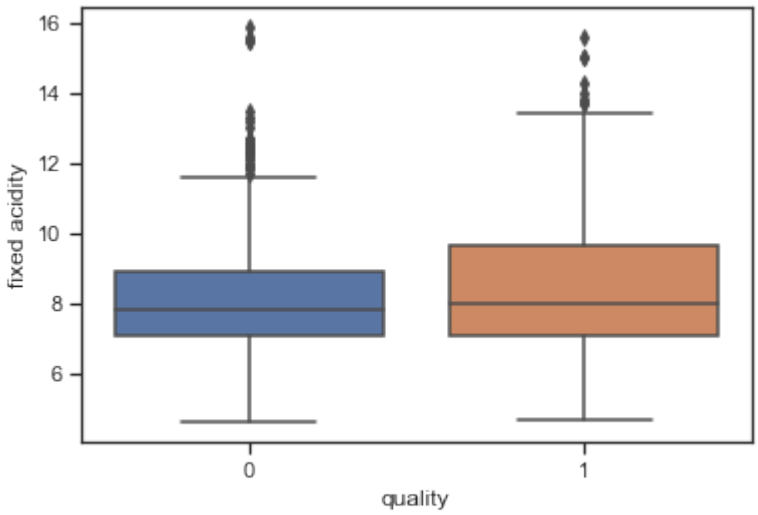
[46]:
```
data['fixed acidity'].plot(xlabel = 'number of sample', ylabel = 'fixed␣
↪acidity')
```

[46]:
```
<AxesSubplot:xlabel='number of sample', ylabel='fixed acidity'>
```



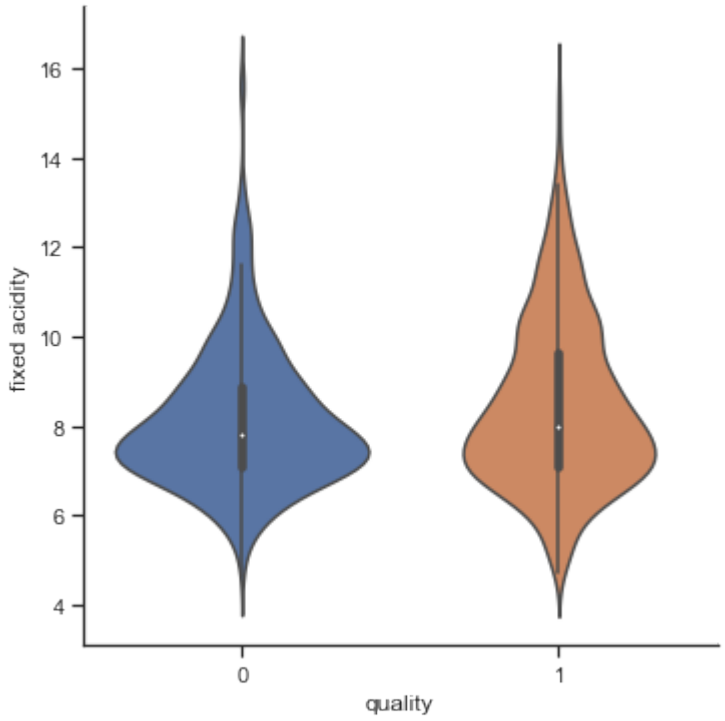
[50]:
```
sns.boxplot(x='quality', y='fixed acidity', data=data)
```

[50]:
```
<AxesSubplot:xlabel='quality', ylabel='fixed acidity'>
```



[51]:
```
sns.catplot(x="quality", y="fixed acidity", data=data, kind="violin")
```

[51]:
```
<seaborn.axisgrid.FacetGrid at 0x7fa31c86b100>
```

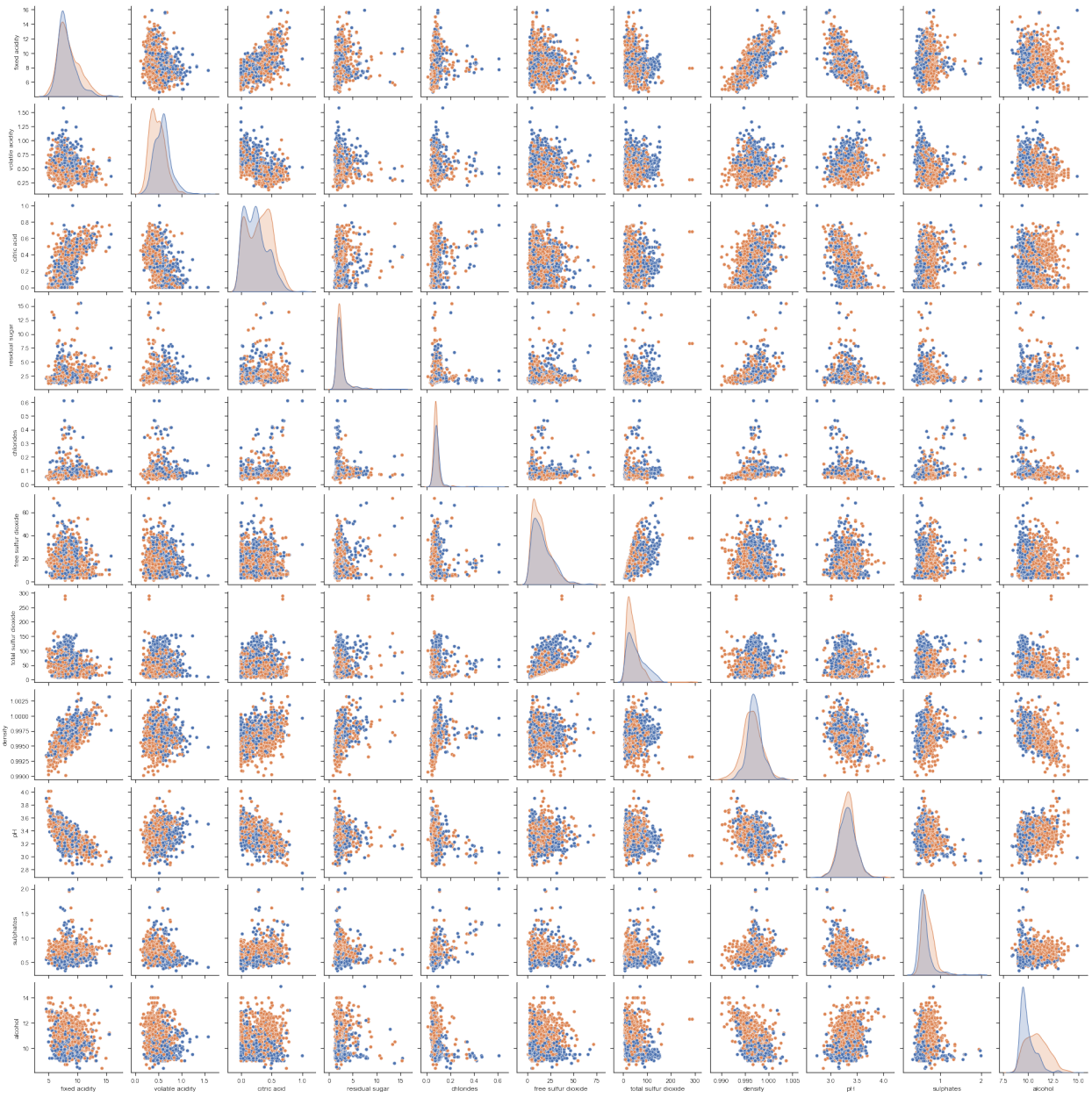


Эти графики показывают плотность распределения, заметны незначительные выбросы с большими значениями кислотности (нарушение нормального распределения)

Рассмотрим матрицу графиков. На пересечении строки и столбца, которые соответстуют двум показателям, строится диаграмма рассеивания. В главной диагонали матрицы строятся гистограммы распределения соответствующих показателей.

[61]:
```
x
```

[61]:
```
<seaborn.axisgrid.PairGrid at 0x7fa3089797c0>
```

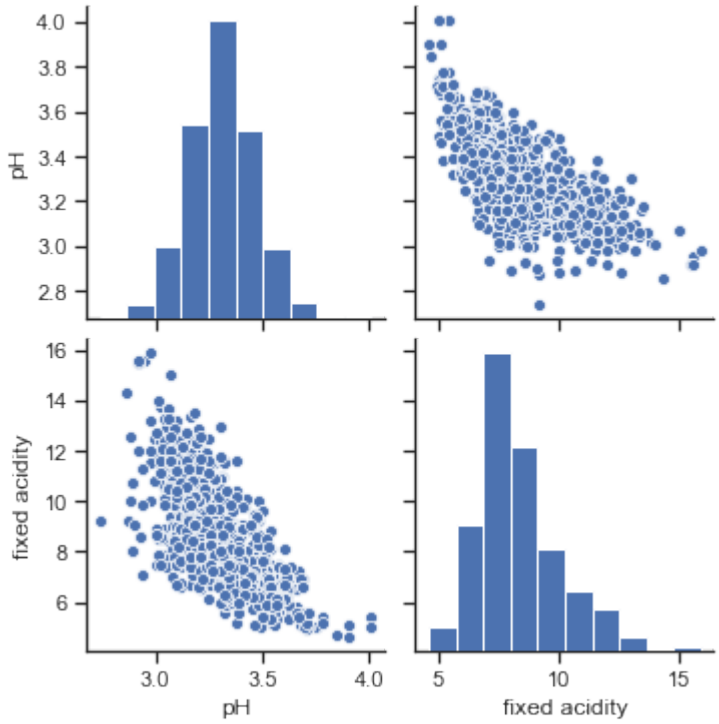


Можем заметить на некоторых графиках что-то похожее на линейность, впоследствии мы увидим это точнее с помощью коэффициентов корреляции.

Рассмотрим подробнее **связь некоторых пар величин**.

[13]:
```
sns.pairplot(data[['pH', 'fixed acidity']])
```

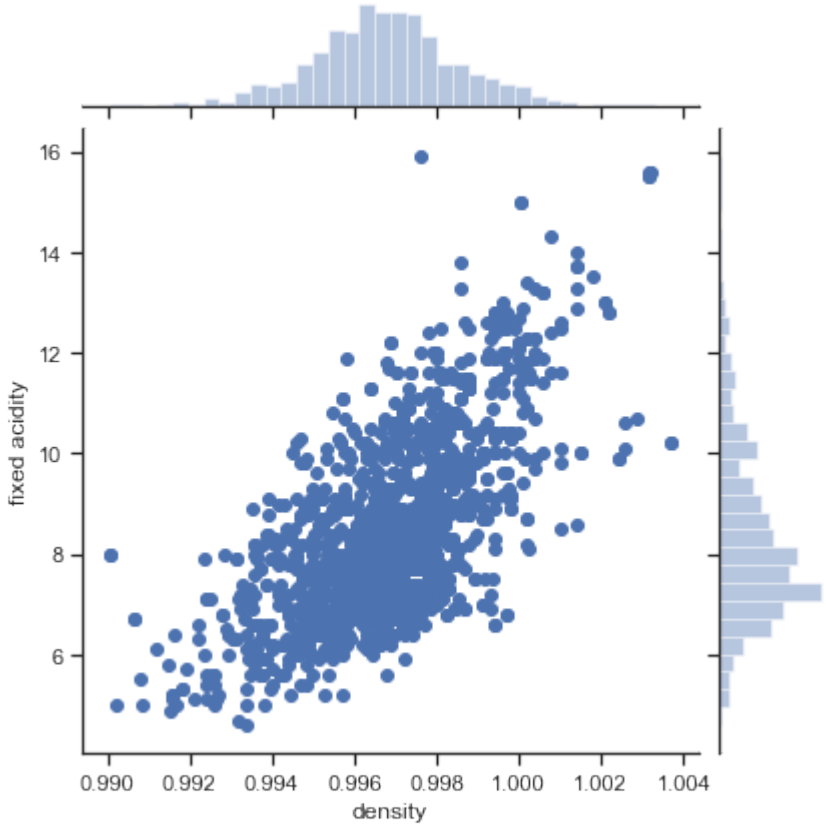
[13]:
```
<seaborn.axisgrid.PairGrid at 0x7fab8f7bd7f0>
```



pH влияет на восприятие кислотности, и возможно поэтому прослеживается какая-то линейная зависимость в измерениях.

[12]:
```
sns.jointplot(x = 'density', y = 'fixed acidity', data=data)
```

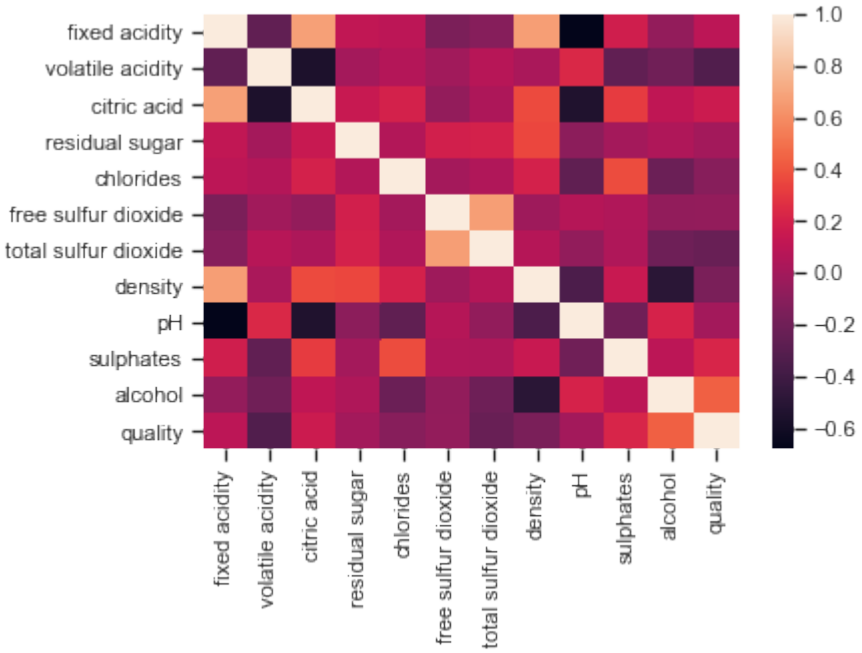
[12]:
```
<seaborn.axisgrid.JointGrid at 0x7fab8e119be0>
```



Прослеживается линейная зависимость фиксированной кислотности и плотности.

[67]:
```
sns.heatmap(data.corr())
```

[67]: <AxesSubplot:>



- Целевой признак наиболее коррелирует с алкогольностью (0,434), летучей кислотностью (-0,321), общим диоксидом серы (-0,232), сульфатами (0,218). Эти признаки оставим точно.
- Целевой признак мало коррелирует с остаточным сахаром (-0,002), pH (-0,003), со свободным диоксидом серы (-0,061), фиксированной кислотностью (0,095). Лучше эти признаки исключить.
- Значительно коррелируют между собой фиксированная кислотность и pH (-0,683). Исключим pH.
- Значительно коррелируют между собой фиксированная кислотность и плотность (-0,668). Исключим фиксированную кислотность.
- Коррелируют voltatile acicidity и citric acid (-0,552). Исключим второй признак как менее коррелирующий с качеством.
- Коррелируют citric acid и pH (-0,542). Исключим второй признак.
- Коррелируют density и alhocol (-0,496). Лучше исключить первый признак.

[19]:
```
df = data.drop(["residual sugar", "pH", "free sulfur dioxide", "fixed␣
↪acidity"], axis = 1)
```