МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Э. Баумана

Факультет «Информатика и системы управления»
Кафедра «Систем обработки информации и управления»

ОТЧЕТ

Лабораторная работа №__1__
по дисциплине «Методы машинного обучения»

Тема: «Создание Истории о данных»

ИСПОЛНИТЕЛЬ: ____Савельев А.А.____
ФИО
группа ИУ5-24М _____
подпись
"__"_____2024 г.

ПРЕПОДАВАТЕЛЬ: ___Гапанюк Ю.Е.__
ФИО
_____
подпись
"__"_____2024 г.

Москва - 2024

Лабораторная работа №1 Создание "истории о данных" (Data Storytelling)

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

Задание:

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь.

Для лабораторных работ не рекомендуется выбирать датасеты очень большого размера.

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.

Выбранный датасет: Student Performance содержит данные о достижениях студентов двух португальских школ в предметах математика и португальский язык.

Он включает следующие столбцы:

1. **school** - школа студента (бинарный: "GP" - Gabriel Pereira или "MS" - Mousinho da Silveira)
2. **sex** - пол студента (бинарный: "F" - женский или "M" - мужской)
3. **age** - возраст студента (числовой: от 15 до 22)
4. **address** - тип домашнего адреса студента (бинарный: "U" - городской или "R" - сельский)
5. **famsize** - размер семьи (бинарный: "LE3" - меньше или равно 3 или "GT3" - больше 3)
6. **Pstatus** - статус совместного проживания родителей (бинарный: "T" - вместе или "A" - раздельно)
7. **Medu** - образование матери (числовой: 0 - нет, 1 - начальное (4-й класс), 2 - 5-9 классы, 3 - среднее, 4 - высшее)

8. **Fedu** - образование отца (числовой: 0 - нет, 1 - начальное (4-й класс), 2 - 5-9 классы, 3 - среднее, 4 - высшее)
9. **Mjob** - профессия матери (номинальный: "teacher", "health", госслужба ("services"), "at_home" или "other")
10. **Fjob** - профессия отца (номинальный: "teacher", "health", госслужба ("services"), "at_home" или "other")
11. **reason** - причина выбора школы (номинальный: близость к "home", "reputation", предпочтение "course" или "other")
12. **guardian** - опекун студента (номинальный: "mother", "father" или "other")
13. **traveltime** - время в пути до школы (числовой: 1 - менее 15 мин., 2 - 15-30 мин., 3 - 30 мин. - 1 час, 4 - более 1 часа)
14. **studytime** - время на учебу в неделю (числовой: 1 - менее 2 часов, 2 - 2-5 часов, 3 - 5-10 часов, 4 - более 10 часов)
15. **failures** - количество прошлых неудач (числовой: n если 1<=n<3, иначе 4)
16. **schoolsup** - дополнительная образовательная поддержка (бинарный: да или нет)
17. **famsup** - семейная образовательная поддержка (бинарный: да или нет)
18. **paid** - дополнительные платные занятия по предмету (математика или португальский) (бинарный: да или нет)
19. **activities** - внеклассные мероприятия (бинарный: да или нет)
20. **nursery** - посещение детского сада (бинарный: да или нет)
21. **higher** - желание получить высшее образование (бинарный: да или нет)
22. **internet** - доступ к интернету дома (бинарный: да или нет)
23. **romantic** - наличие романтических отношений (бинарный: да или нет)
24. **famrel** - качество семейных отношений (числовой: от 1 - очень плохие до 5 - отличные)
25. **freetime** - свободное время после школы (числовой: от 1 - очень мало до 5 - очень много)
26. **goout** - прогулки с друзьями (числовой: от 1 - очень редко до 5 - очень часто)
27. **Dalc** - потребление алкоголя в будние дни (числовой: от 1 - очень низкое до 5 - очень высокое)
28. **Walc** - потребление алкоголя в выходные (числовой: от 1 - очень низкое до 5 - очень высокое)
29. **health** - текущее состояние здоровья (числовой: от 1 - очень плохое до 5 - очень хорошее)
30. **absences** - количество пропусков занятий (числовой: от 0 до 93)
31. **G1** - оценка за первый период (числовой: от 0 до 20)
32. **G2** - оценка за второй период (числовой: от 0 до 20)
33. **G3** - итоговая оценка (числовой: от 0 до 20, целевая переменная)

```
# Импорт необходимых библиотек
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
%matplotlib inline
sns.set(style="ticks")

# Подгрузка набора данных
data = pd.read_csv(r'student/student-mat.csv', sep=";")

# Вывод содержания датасета
data.head(15).T
```

```
                 0        1        2         3       4           5
6   \
school          GP       GP       GP        GP      GP          GP
GP
sex              F        F        F         F       F           M
M
age             18       17       15        15      16          16
16
address          U        U        U         U       U           U
U
famsize        GT3      GT3      LE3       GT3     GT3         LE3
LE3
Pstatus          A        T        T         T       T           T
T
Medu             4        1        1         4       3           4
2
Fedu             4        1        1         2       3           3
2
Mjob        at_home  at_home  at_home    health   other    services
other
Fjob        teacher    other    other  services   other       other
other
reason       course   course    other      home    home  reputation
home
guardian     mother   father   mother    mother  father      mother
mother
traveltime        2        1        1         1       1           1
1
studytime         2        2        2         3       2           2
2
failures          0        0        3         0       0           0
0
schoolsup       yes       no      yes        no      no          no
no
famsup           no      yes       no       yes     yes         yes
no
paid             no       no      yes       yes     yes         yes
no
activities       no       no       no       yes      no         yes
no
nursery         yes       no      yes       yes     yes         yes
```

```
yes
higher           yes       yes       yes        yes      yes          yes
yes
internet          no       yes       yes        yes       no          yes
yes
romantic          no        no        no        yes       no           no
no
famrel             4         5         4          3        4            5
4
freetime           3         3         3          2        3            4
4
goout              4         3         2          2        2            2
4
Dalc               1         1         2          1        1            1
1
Walc               1         1         3          1        2            2
1
health             3         3         3          5        5            5
3
absences           6         4        10          2        4           10
0
G1                 5         5         7         15        6           15
12
G2                 6         5         8         14       10           15
12
G3                 6         6        10         15       10           15
11

                   7         8         9          10          11
12  \
school            GP        GP        GP          GP          GP
GP
sex                F         M         M           F           F
M
age               17        15        15          15          15
15
address            U         U         U           U           U
U
famsize          GT3       LE3       GT3         GT3         GT3
LE3
Pstatus            A         A         T           T           T
T
Medu               4         3         3           4           2
4
Fedu               4         2         4           4           1
4
Mjob           other  services     other     teacher    services
health
Fjob        teacher     other     other      health       other
```

```
services
reason          home      home    home  reputation  reputation
course
guardian      mother    mother  mother      mother      father
father
traveltime         2         1       1           1           3
1
studytime          2         2       2           2           3
1
failures           0         0       0           0           0
0
schoolsup        yes        no      no          no          no
no
famsup           yes       yes     yes         yes         yes
yes
paid              no       yes     yes         yes          no
yes
activities        no        no     yes          no         yes
yes
nursery          yes       yes     yes         yes         yes
yes
higher           yes       yes     yes         yes         yes
yes
internet          no       yes     yes         yes         yes
yes
romantic          no        no      no          no          no
no
famrel             4         4       5           3           5
4
freetime           1         2       5           3           2
3
goout              4         2       1           3           2
3
Dalc               1         1       1           1           1
1
Walc               1         1       1           2           1
3
health             1         1       5           2           4
5
absences           6         0       0           0           4
2
G1                 6        16      14          10          10
14
G2                 5        18      15           8          12
14
G3                 6        19      15           9          12
14

                  13     14
```

```
school           GP     GP
sex               M      M
age              15     15
address           U      U
famsize         GT3    GT3
Pstatus           T      A
Medu              4      2
Fedu              3      2
Mjob        teacher  other
Fjob          other  other
reason       course   home
guardian     mother  other
traveltime        2      1
studytime         2      3
failures          0      0
schoolsup        no     no
famsup          yes    yes
paid            yes     no
activities       no     no
nursery         yes    yes
higher          yes    yes
internet        yes    yes
romantic         no    yes
famrel            5      4
freetime          4      5
goout             3      2
Dalc              1      1
Walc              2      1
health            3      3
absences          2      0
G1               10     14
G2               10     16
G3               11     16
```

```
# Размерность датасета
data.shape
```

```
(395, 33)
```

```
# Колонки
data.columns
```

```
Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus',
'Medu', 'Fedu',
       'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime',
'studytime',
       'failures', 'schoolsup', 'famsup', 'paid', 'activities',
'nursery',
       'higher', 'internet', 'romantic', 'famrel', 'freetime',
'goout', 'Dalc',
```

```
       'Walc', 'health', 'absences', 'G1', 'G2', 'G3'],
      dtype='object')
```

```
# Типы данных колонок
data.dtypes
```

```
school        object
sex           object
age            int64
address       object
famsize       object
Pstatus       object
Medu           int64
Fedu           int64
Mjob          object
Fjob          object
reason        object
guardian      object
traveltime     int64
studytime      int64
failures       int64
schoolsup     object
famsup        object
paid          object
activities    object
nursery       object
higher        object
internet      object
romantic      object
famrel         int64
freetime       int64
goout          int64
Dalc           int64
Walc           int64
health         int64
absences       int64
G1             int64
G2             int64
G3             int64
dtype: object
```

Проверим набор данных на наличие пустых значений

```
flag = 0
# 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]
```

```
    if temp_null_count == 0:
        continue
    else:
        flag = 1
        print('{} - {}'.format(col, temp_null_count))
if flag == 0:
    print("Пустых значений нет")
```

```
Пустых значений нет
```

```
# Основные статистические характеристки набора данных
data.describe()
```

```
              age        Medu        Fedu  traveltime   studytime
failures  \
count  395.000000  395.000000  395.000000  395.000000  395.000000
395.000000
mean    16.696203    2.749367    2.521519    1.448101    2.035443
0.334177
std      1.276043    1.094735    1.088201    0.697505    0.839240
0.743651
min     15.000000    0.000000    0.000000    1.000000    1.000000
0.000000
25%     16.000000    2.000000    2.000000    1.000000    1.000000
0.000000
50%     17.000000    3.000000    2.000000    1.000000    2.000000
0.000000
75%     18.000000    4.000000    3.000000    2.000000    2.000000
0.000000
max     22.000000    4.000000    4.000000    4.000000    4.000000
3.000000

           famrel    freetime       goout        Dalc        Walc
health  \
count  395.000000  395.000000  395.000000  395.000000  395.000000
395.000000
mean     3.944304    3.235443    3.108861    1.481013    2.291139
3.554430
std      0.896659    0.998862    1.113278    0.890741    1.287897
1.390303
min      1.000000    1.000000    1.000000    1.000000    1.000000
1.000000
25%      4.000000    3.000000    2.000000    1.000000    1.000000
3.000000
50%      4.000000    3.000000    3.000000    1.000000    2.000000
4.000000
75%      5.000000    4.000000    4.000000    2.000000    3.000000
5.000000
max      5.000000    5.000000    5.000000    5.000000    5.000000
5.000000
```

	absences	G1	G2	G3
count	395.000000	395.000000	395.000000	395.000000
mean	5.708861	10.908861	10.713924	10.415190
std	8.003096	3.319195	3.761505	4.581443
min	0.000000	3.000000	0.000000	0.000000
25%	0.000000	8.000000	9.000000	8.000000
50%	4.000000	11.000000	11.000000	11.000000
75%	8.000000	13.000000	13.000000	14.000000
max	75.000000	19.000000	19.000000	20.000000

Визуальный анализ набора данных

Гистограмма

Позволяет оценить плотность вероятности распределения данных. Выберем несколько численных параметров и сделаем для них гистограмму:

- studytime - время на учебу в неделю (числовой: 1 - менее 2 часов, 2 - 2-5 часов, 3 - 5-10 часов, 4 - более 10 часов)
- age - возраст студента (числовой: от 15 до 22)
- G1 - оценка за первый период (числовой: от 0 до 20)
- G3 - итоговая оценка (числовой: от 0 до 20, целевая переменная)

```
# Колонки
columns = ['studytime', 'age', 'G1', 'G3']

fig, axes = plt.subplots(1, len(columns), figsize=(20, 5))

# Рассчет гистограммы по колонкам
for i, col in enumerate(columns):
    sns.histplot(data[col], ax=axes[i], kde=True, stat='density')
    axes[i].set_title(f'Распределение {col}')
    axes[i].set_ylabel('Плотность')

plt.tight_layout()
plt.show()
```

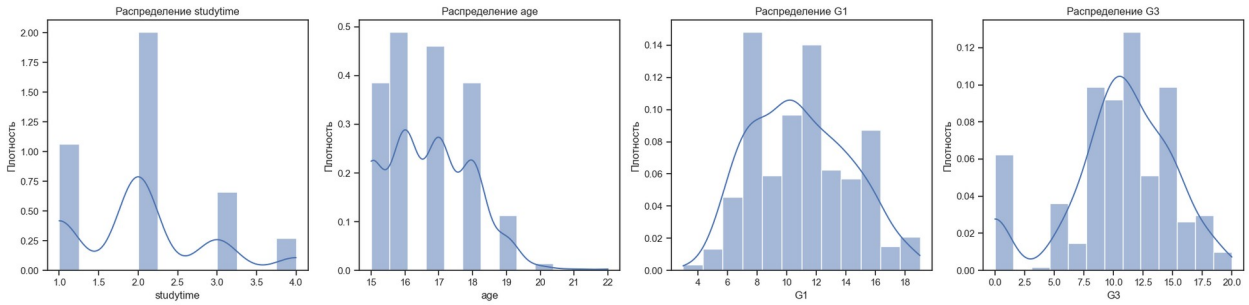
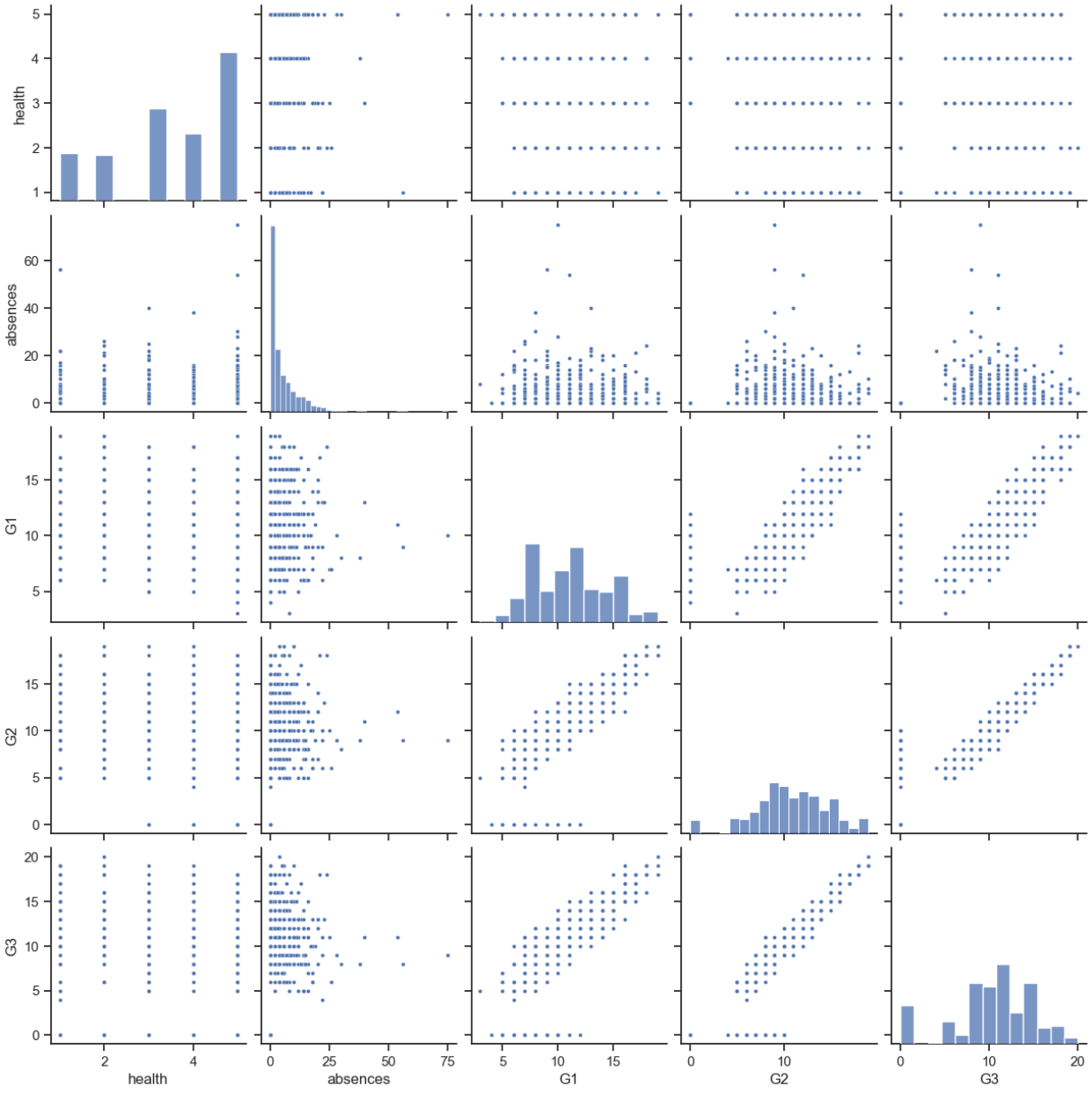


Диаграмма разброса

```
# Оставим некоторые численные признаки
columns_to_plot = ['health', 'absences', 'G1', 'G2', 'G3']
numeric_data = data[columns_to_plot]

# Построение pairplot
sns.pairplot(numeric_data, plot_kws={'s': 10})
plt.show()
```



По полученным диаграммам можно сделать выводы:

Взаимосвязь оценок:

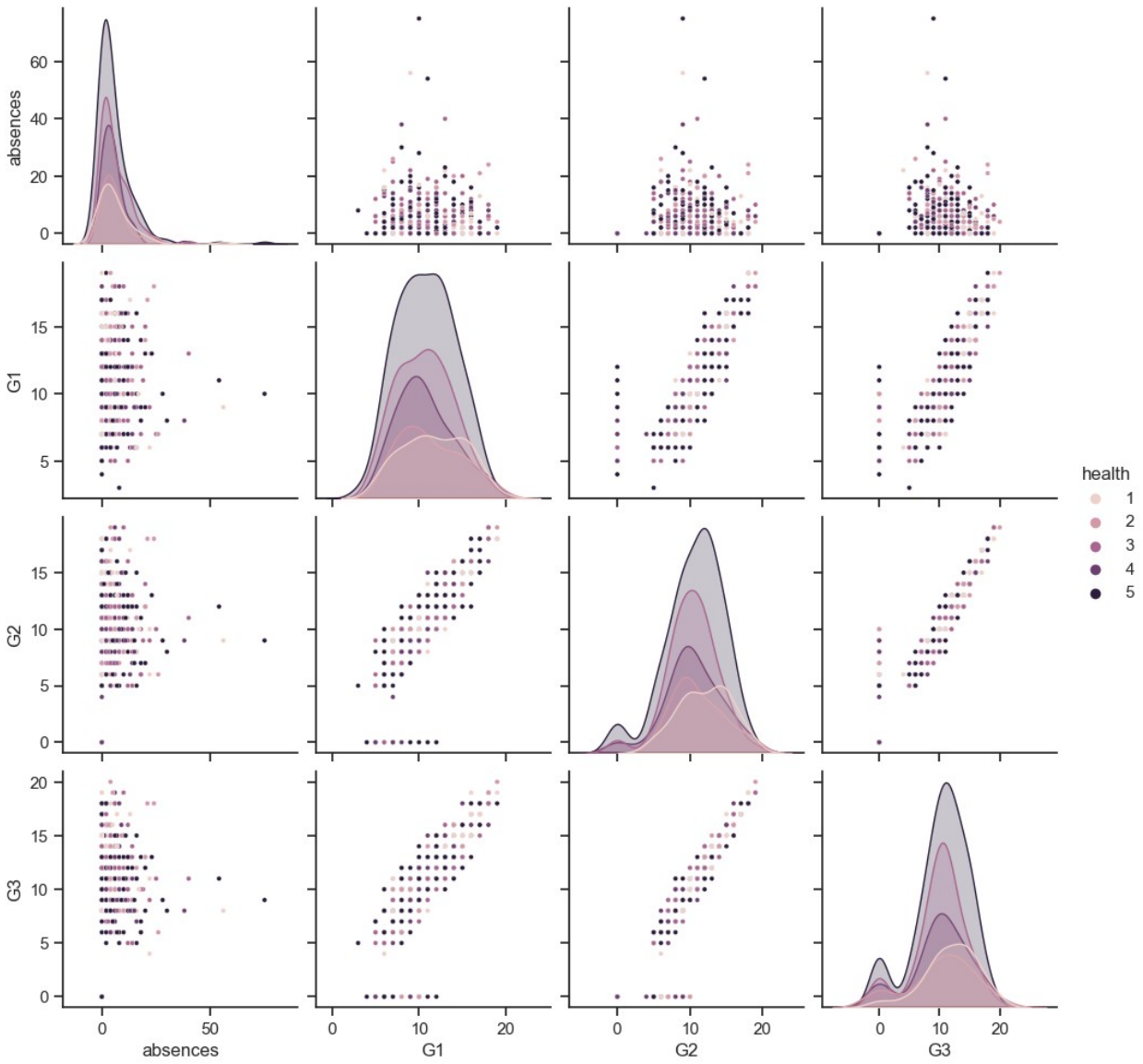
- Оценки за первый (G1), второй (G2) и третий (G3) периоды имеют положительную корреляцию, что ожидаемо, так как они связаны с общей успеваемостью студента.
- На диагональных гистограммах видно распределение оценок, где большинство студентов получают оценки в диапазоне от 5 до 15.

Пропуски и здоровье:

- Пропуски занятий (absences) также не показывают явной зависимости от состояния здоровья (health).

```
# Оставим некоторые численные признаки
columns_to_plot = ['health', 'absences', 'G1', 'G2', 'G3']
numeric_data = data[columns_to_plot]

# Построение pairplot
sns.pairplot(numeric_data, plot_kws={'s': 10}, hue='health')
plt.show()
```



На данной диаграмме разброса можно оценить разброс оценок и пропусков в зависимости от состояния здоровья:

- так видно, что плохое самочувствие увеличивает распределение оценок, захватывая тем самым и плохие
- как нестранно плохое самочувствие не сильно походит на причину большого количества пропусков занятий

Boxplot

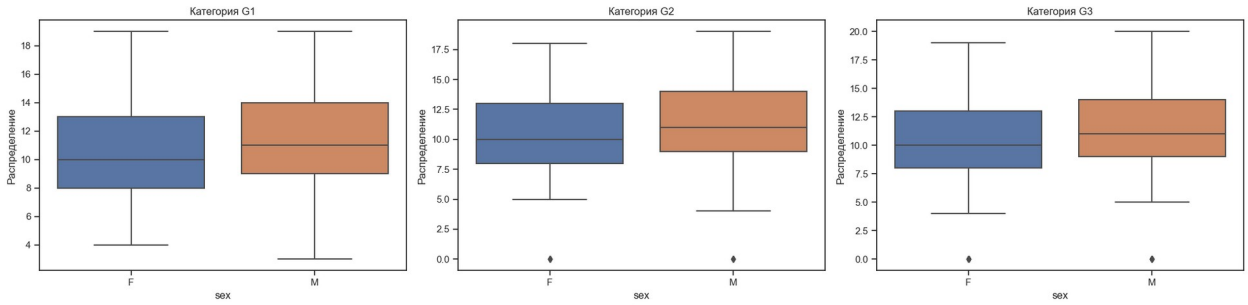
Используем этот тип графика для визуализации распределения оценок по категориям

```
# Колонки
columns = ['G1', 'G2', 'G3']
```

```
fig, axes = plt.subplots(1, len(columns), figsize=(20, 5))

# Рассчет гистограммы по колонкам
for i, col in enumerate(columns):
    sns.boxplot(x='sex', y=col, data=data, ax=axes[i])
    axes[i].set_title(f'Категория {col}')
    axes[i].set_ylabel('Распределение')

plt.tight_layout()
plt.show()
```



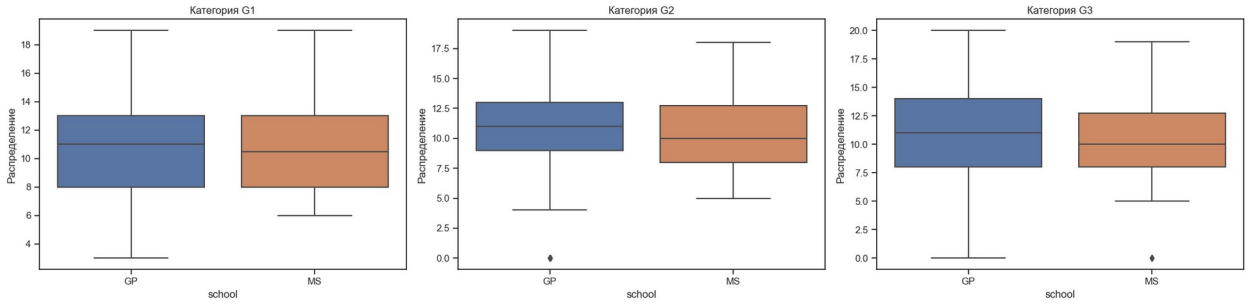
По данным графикам можно сказать что величина и распределение оценки у мальчиков выше чем у девочек

```
# Колонки
columns = ['G1', 'G2', 'G3']

fig, axes = plt.subplots(1, len(columns), figsize=(20, 5))

# Рассчет гистограммы по колонкам
for i, col in enumerate(columns):
    sns.boxplot(x='school', y=col, data=data, ax=axes[i])
    axes[i].set_title(f'Категория {col}')
    axes[i].set_ylabel('Распределение')

plt.tight_layout()
plt.show()
```

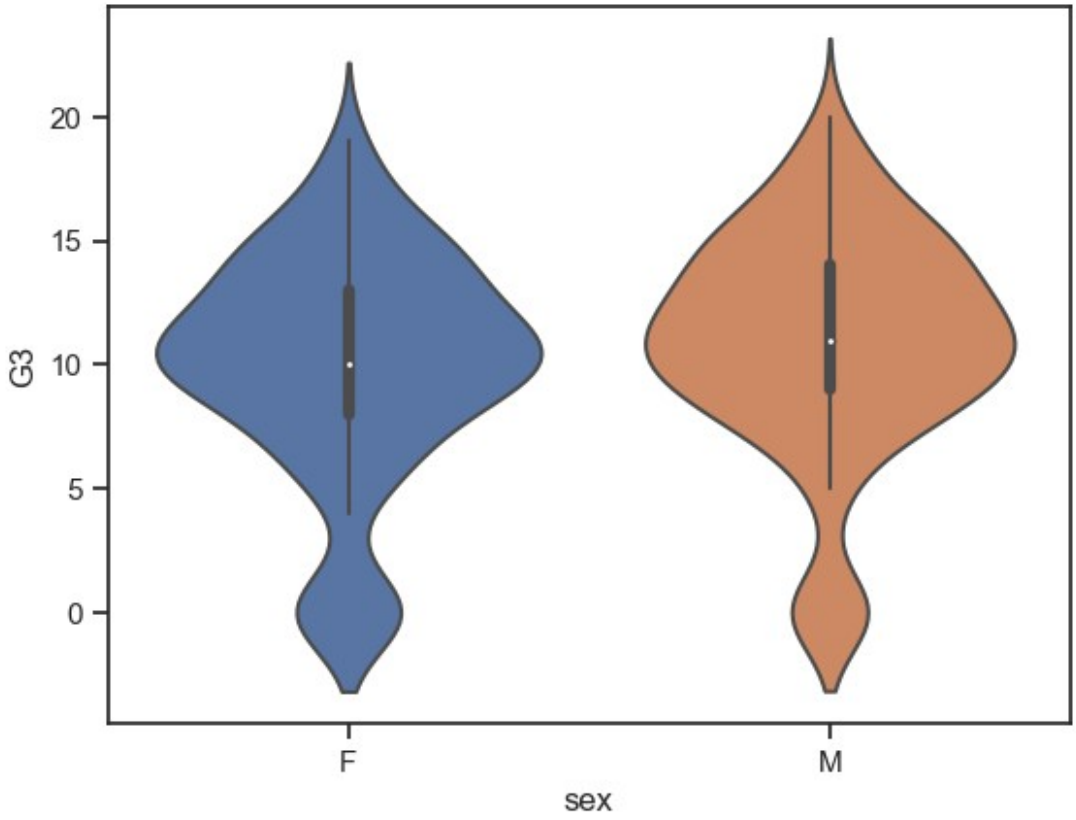


На этих графиках можно оценить распределение оценок по школам

Violinplot

Хорошо показывает распределение данных и их плотность по категориям

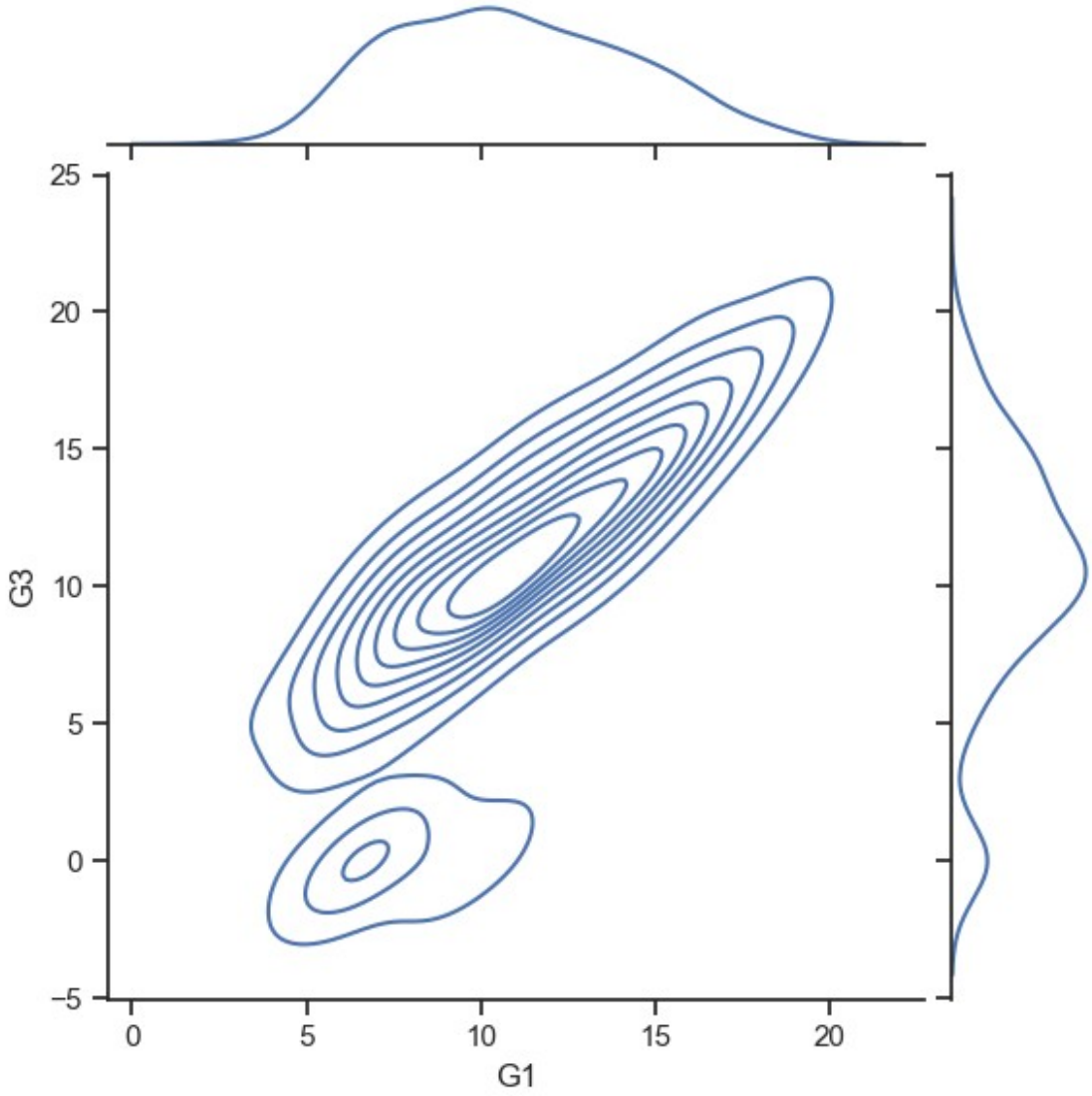
```
sns.violinplot(x='sex', y='G3', data=data)
plt.show()
```



Jointplot

Подходит для отображения взаимосвязи между двумя числовыми признаками

```
sns.jointplot(x='G1', y='G3', data=data, kind='kde')
plt.show()
```



Выводы:

Корреляция между оценками:

- Видно, что оценки за первый период (G1) и итоговые оценки (G3) имеют положительную корреляцию. Чем выше оценка за первый период, тем выше итоговая оценка.

Группы студентов:

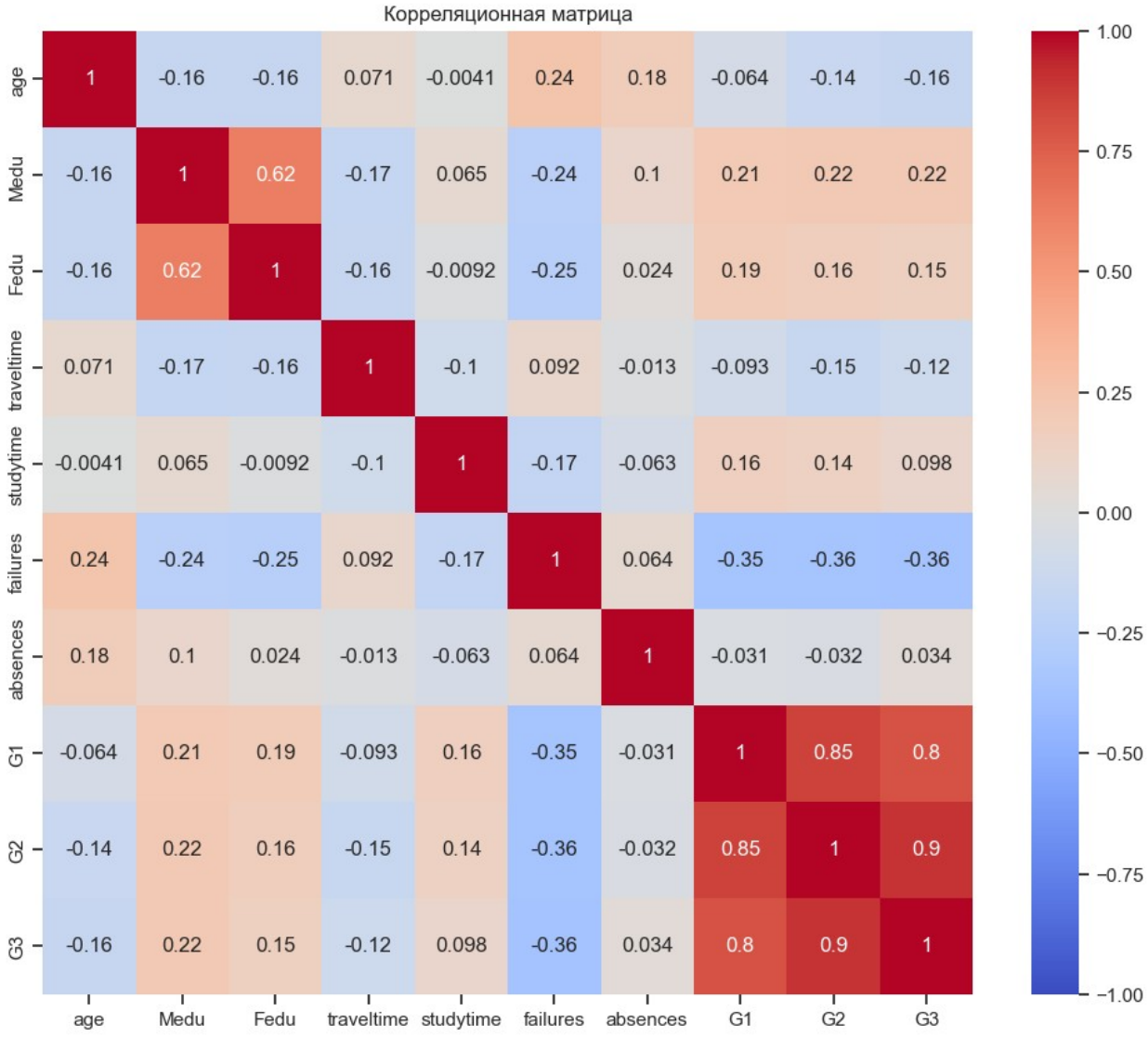
- Наблюдаются две основные группы студентов: одна с высокими оценками (около 10-15 баллов) и другая с низкими оценками (около 0-5 баллов).

Корреляционная матрица

```
columns_to_plot = ['age', 'Medu', 'Fedu', 'traveltime', 'studytime',
'failures', 'absences', 'G1', 'G2', 'G3']
```

```
correlation_matrix = data[columns_to_plot].corr()

# Построение корреляционной матрицы
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1,
vmax=1)
plt.title('Корреляционная матрица')
plt.show()
```



Выводы по корреляционной матрице:

Корреляция между оценками:

- Оценки за первый (G1), второй (G2) и третий (G3) периоды имеют высокую положительную корреляцию (0.85-0.9), что указывает на сильную взаимосвязь между этими оценками.

Образование родителей:

- Образование матери (Medu) и отца (Fedu) имеют значительную положительную корреляцию (0.62), что указывает на то, что уровень образования родителей часто совпадает.

Неудачи и оценки:

- Количество неудач (failures) имеет отрицательную корреляцию с оценками (G1, G2, G3), что достаточно логично, поскольку большее количество неудач обычно ведет к более низким итоговым оценкам.

Возраст и пропуски:

- Пропуски (absences) имеют слабую положительную корреляцию с возрастом (0.18), что может указывать на тенденцию более старших студентов пропускать больше занятий.

Время на учебу и оценки:

- Время на учебу (studytime) имеет слабую положительную корреляцию с итоговой оценкой (G3), что подтверждает, что большее время на учебу может приводить к лучшим оценкам.

Выводы

Создание «истории о данных» позволяет провести визуальный разведочный анализ датасета, не производя комплексных вычислений для оценки его основных характеристик и пригодности в машинном обучений.