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

Лабораторная работа №1
по дисциплине
«Методы машинного обучения»
на тему

«Разведочный анализ данных. Исследование и визуализация данных»

Выполнил:
студент группы ИУ5-21М
Наинг Ко Ко Линн

Москва — 2020 г.

1. Цель лабораторной работы

Изучить различные методы визуализации данных [1].

2. Задание

Требуется выполнить следующие действия [1]:

- Выбрать набор данных (датасет).
- Создать ноутбук, который содержит следующие разделы:
    1. Текстовое описание выбранного набора данных.
    2. Основные характеристики датасета.
    3. Визуальное исследование датасета.
    4. Информация о корреляции признаков.
- Сформировать отчет и разместить его в своем репозитории на GitHub

3. Ход выполнения работы

3.1. Текстовое описание набора данных

Сердце - это удивительный орган. Он бьется ровно, ровно, примерно от 60 до 100 раз каждую минуту. Это примерно 100 000 раз в день. Иногда твое сердце выходит из ритма. Ваш врач называет нерегулярное или неправильное сердцебиение аритмией. Аритмия (также называемая дисритмией) может вызывать неравномерное сердцебиение или сердцебиение, которое либо слишком медленное, либо слишком быстрое.

3.2. Основные характеристики набора данных

Подключим все необходимые библиотеки [1]

```
[ ]  from datetime import datetime
     import pandas as pd
     import seaborn as sns
```

```
[ ]  # Enable inline plots
     %matplotlib inline
     # Set plot style
     sns.set(style="ticks")
     # Set plots formats to save high resolution PNG
     from IPython.display import set_matplotlib_formats
     set_matplotlib_formats("retina")
```

```
[30] pd.set_option("display.width", 70)
```

```
     data = pd.read_csv("./heart.csv")
```

Настроим отображение графиков [3,4]:

Загрузим непосредственно данные[5]

```
In [0]: data.dtypes
```

```
Out[0]: 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
```

Посмотрим на данные в данном наборе данных:

In [0]:
```
data.head()
```

Out[0]:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	fai
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0
1	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0
2	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3
3	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0
4	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0

Проверим размер набора данных:

In [0]:
```
df=data.copy()
df.shape
```

Out[0]: (395, 33)

Проверим основные статистические характеристики набора данных:

In [0]:
```
df.describe()
```

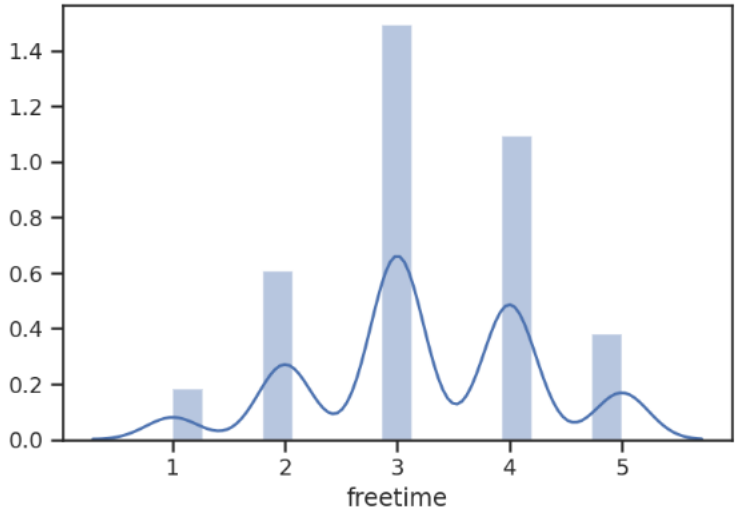
Out[0]:

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.0
mean	16.696203	2.749367	2.521519	1.448101	2.035443	0.334177	3.944304	3.235443	3.108861	1.48
std	1.276043	1.094735	1.088201	0.697505	0.839240	0.743651	0.896659	0.998862	1.113278	0.89
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.00
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000	4.000000	3.000000	2.000000	1.00
50%	17.000000	3.000000	2.000000	1.000000	2.000000	0.000000	4.000000	3.000000	3.000000	1.00
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	4.000000	4.000000	2.00
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	5.000000	5.000000	5.00

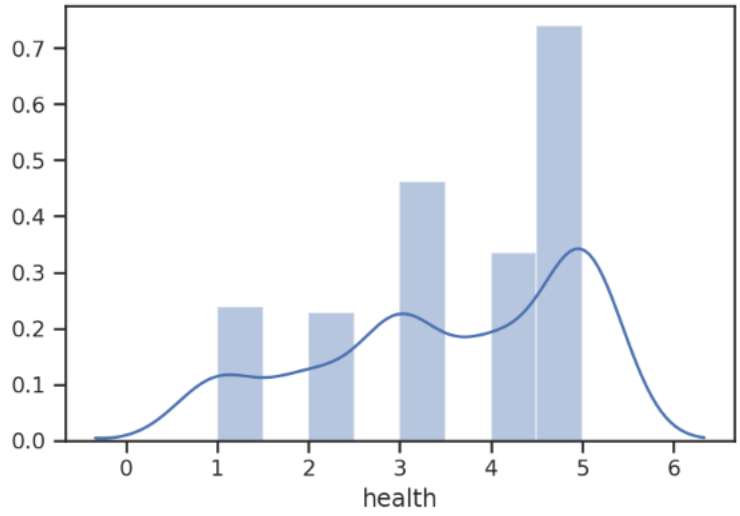
3.3. Визуальное исследование датасета

Давайте оценим распределение целевого атрибута - Рейтинг:

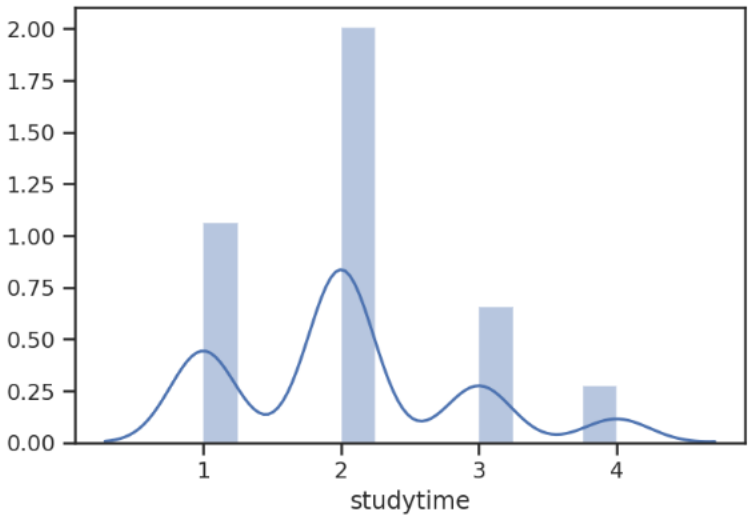
In [0]:
```
sns.distplot(df["freetime"]);
```



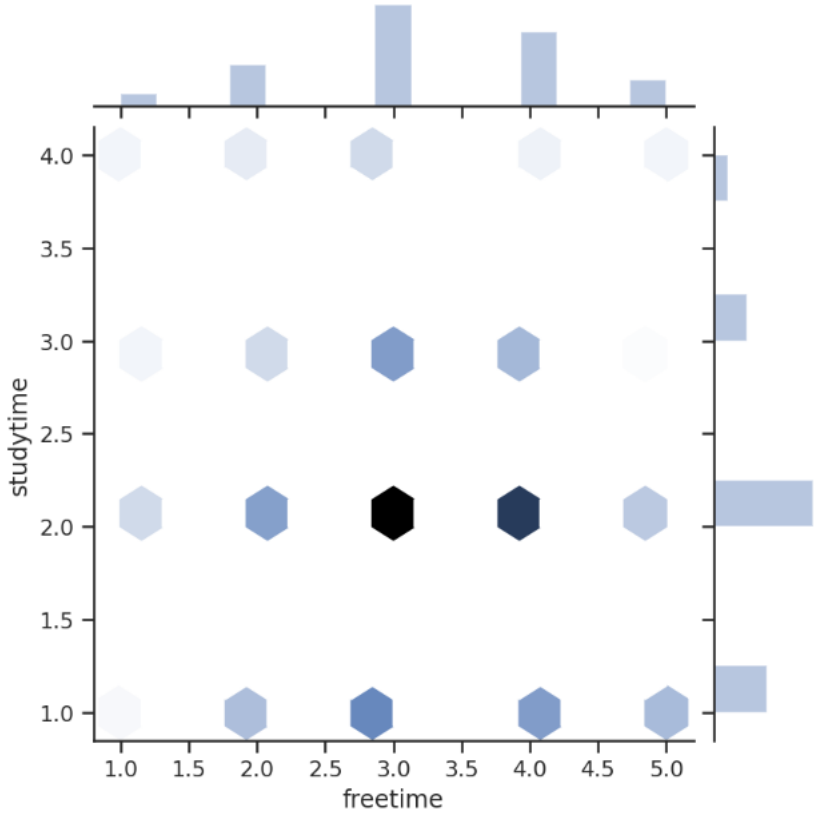
In [0]:
```
sns.distplot(df["health"]);
```



In [0]:
```
sns.distplot(df["studytime"]);
```



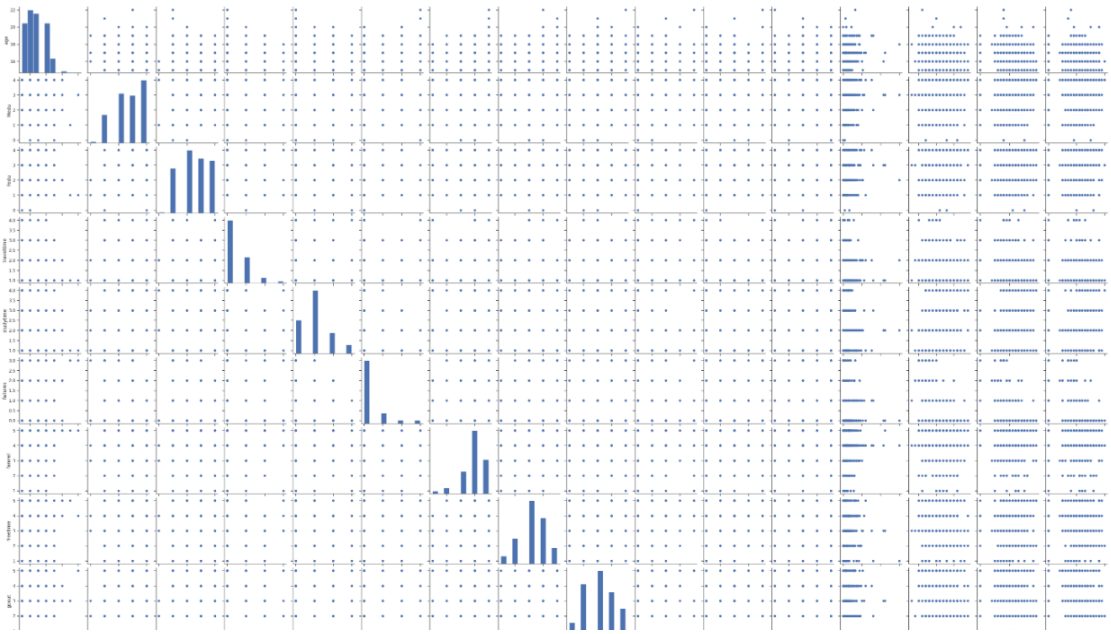
In [0]:
```
sns.jointplot(x="freetime", y="studytime", data=df, kind="hex");
```



Построим парные диаграммы по всем показателям по исходному набору данных:

```
sns.pairplot(df, plot_kws=dict(linewidth=0));
```

In [0]:
```
sns.pairplot(df, plot_kws=dict(linewidth=0));
```



3.4. Информация о корреляции признаков

In [0]:
```
df.corr()
```

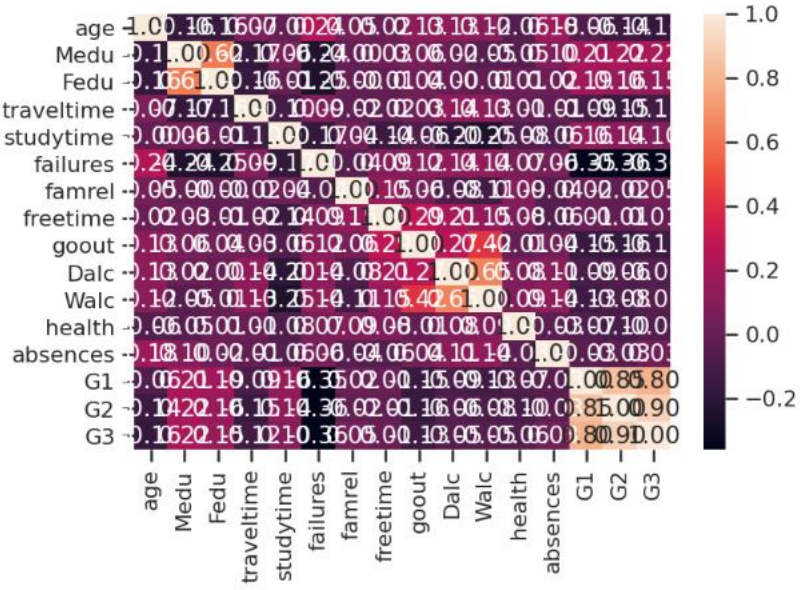
Out[0]:

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Wa
age	1.000000	-0.163658	-0.163438	0.070641	-0.004140	0.243665	0.053940	0.016434	0.126964	0.131125	0.1
Medu	-0.163658	1.000000	0.623455	-0.171639	0.064944	-0.236680	-0.003914	0.030891	0.064094	0.019834	-0.0
Fedu	-0.163438	0.623455	1.000000	-0.158194	-0.009175	-0.250408	-0.001370	-0.012846	0.043105	0.002386	-0.0
traveltime	0.070641	-0.171639	-0.158194	1.000000	-0.100909	0.092239	-0.016808	-0.017025	0.028540	0.138325	0.1
studytime	-0.004140	0.064944	-0.009175	-0.100909	1.000000	-0.173563	0.039731	-0.143198	-0.063904	-0.196019	-0.2
failures	0.243665	-0.236680	-0.250408	0.092239	-0.173563	1.000000	-0.044337	0.091987	0.124561	0.136047	0.1
famrel	0.053940	-0.003914	-0.001370	-0.016808	0.039731	-0.044337	1.000000	0.150701	0.064568	-0.077594	-0.1
freetime	0.016434	0.030891	-0.012846	-0.017025	-0.143198	0.091987	0.150701	1.000000	0.285019	0.209001	0.1
goout	0.126964	0.064094	0.043105	0.028540	-0.063904	0.124561	0.064568	0.285019	1.000000	0.266994	0.4
Dalc	0.131125	0.019834	0.002386	0.138325	-0.196019	0.136047	-0.077594	0.209001	0.266994	1.000000	0.6
Walc	0.117276	-0.047123	-0.012631	0.134116	-0.253785	0.141962	-0.113397	0.147822	0.420386	0.647544	1.0
health	-0.062187	-0.046878	0.014742	0.007501	-0.075616	0.065827	0.094056	0.075733	-0.009577	0.077180	0.0
absences	0.175230	0.100285	0.024473	-0.012944	-0.062700	0.063726	-0.044354	-0.058078	0.044302	0.111908	0.1
G1	-0.064081	0.205341	0.190270	-0.093040	0.160612	-0.354718	0.022168	0.012613	-0.149104	-0.094159	-0.1

Построим корреляционную матрицу по всему набору данных:

Визуализируем корреляционную матрицу с помощью тепловой карты:

```
In [0]: sns.heatmap(df.corr(), annot=True, fmt=".2f");
```



Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2019. — Режим доступа: https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION (дата обращения: 13.02.2019)

[2] https://www.kaggle.com/datasets