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
ЧжаоЛян

Москва — 2022 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

В этой тетради я буду использовать графики для визуализации взаимосвязи между переменными в наборе данных "Успеваемость студентов на экзаменах".

- **The dataset includes the following columns:**
  **Gender** (object) - The gender of the participant
  - female
  - male
- **Race/Ethnicity** (object) - The race or ethnicity of the participant
  - Group A
  - Group B
  - Group C
  - Group D
  - Group E
- **Parental level of education** (object) - The parental level of education of the participant
  - bachelor's degree
  - some college
  - master's degree
  - associate's degree
  - high school
  - some high school
- **Lunch** (object) - Whether the participant is:

    - standard
    - free/ reduced
- **Test preparation course** (object) - Whether the participant took the test preparation course or not
    - none
    - completed
- **Math score** (int64) - The participants math score
- **Reading score** (int64) - The participants reading score
- **Writing score** (int64) - The participants writing score

## 3.2. Основные характеристики набора данных

Подключим все необходимые библиотеки & Загрузим непосредственно данные:

In [1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

sns.set()
sns.set_palette('Set2')
```

Изучение данных необходимо для того, чтобы узнать, с какими данными вы имеете дело, и решить, какие графики будут применимы в каждой переменной.

In [2]:
```
df = pd.read_csv('../input/students-performance-in-exams/StudentsPerformance.csv')
df.head()
```

Out[2]:

| | gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score |
|---|---|---|---|---|---|---|---|---|
| 0 | female | group B | bachelor's degree | standard | none | 72 | 72 | 74 |
| 1 | female | group C | some college | standard | completed | 69 | 90 | 88 |
| 2 | female | group B | master's degree | standard | none | 90 | 95 | 93 |
| 3 | male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 |
| 4 | male | group C | some college | standard | none | 76 | 78 | 75 |

In [3]:
```
df.info()
```

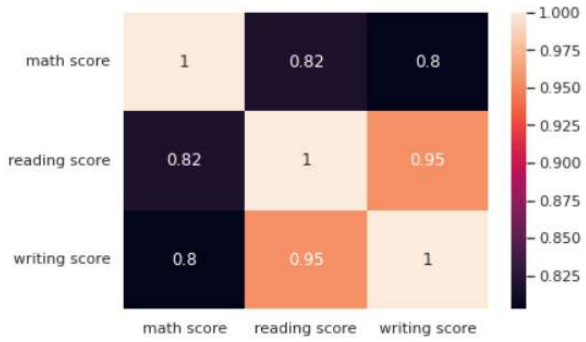
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental level of education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test preparation course      1000 non-null   object
 5   math score                   1000 non-null   int64 
 6   reading score                1000 non-null   int64 
 7   writing score                1000 non-null   int64 
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

## Баллы Корреляции

Давайте проверим корреляцию между нашими числовыми переменными, которыми являются баллы по математике, чтению и письму.

In [4]:
```
df_corr = df.corr()
sns.heatmap(df_corr, annot=True)
```

Out[4]:
```
<AxesSubplot:>
```




На этой тепловой карте корреляции легко заметить, что оценки за чтение и письмо очень сильно коррелируют друг с другом, а оценки по математике

также коррелируют с другими оценками, но не так сильно, как корреляция оценок за чтение и письмо.

Используя диаграммы рассеяния, мы можем лучше представить корреляцию между тремя числовыми переменными.

```
In [5]:
fig, ax = plt.subplots(ncols=3, figsize=(20,7))

fig.suptitle('Score Distributions of Students')

sns.scatterplot(data=df, x='math score', y='writing score', ax=ax[0], alpha=0.6)
sns.scatterplot(data=df, x='math score', y='reading score', ax=ax[1], alpha=0.6)
sns.scatterplot(data=df, x='writing score', y='reading score', ax=ax[2], alpha=0.6)
```

```
Out[5]:
<AxesSubplot:xlabel='writing score', ylabel='reading score'>
```

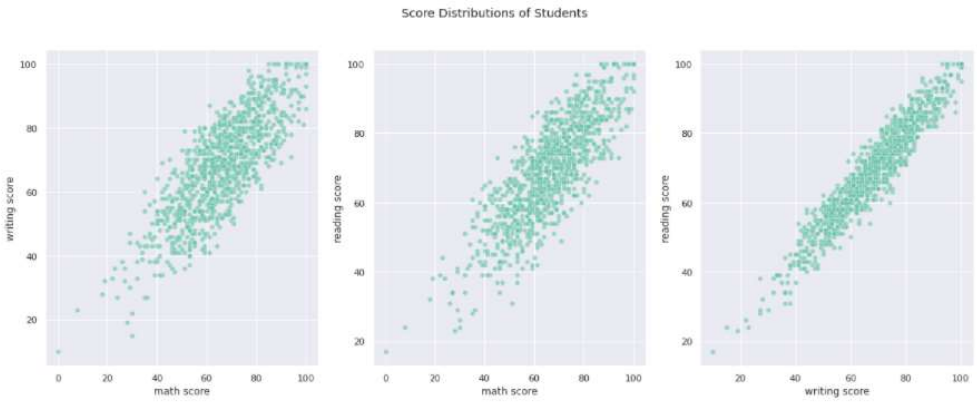



График показывает, что переменные имеют положительную корреляцию, хотя графики оценок за чтение и письмо расположены ближе друг к другу, чем другие графики, включающие оценку за математику

## Курс подготовки к тестированию

Добавим еще одну переменную, чтобы получить больше информации, в данном случае мы добавим "курс подготовки к тесту", который означает,

проходил ли студент курс подготовки к тесту.

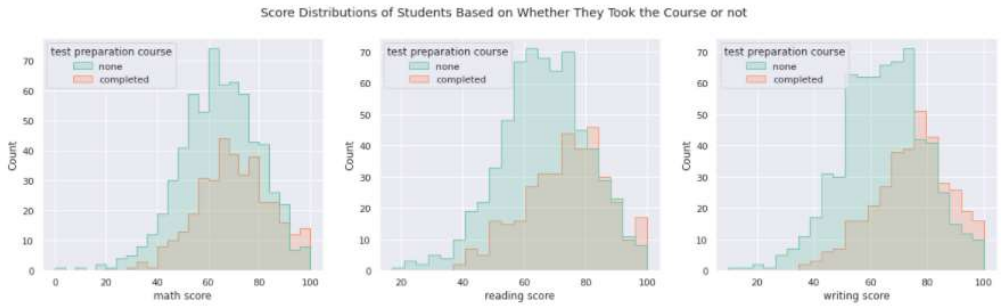
In [6]:

```
fig, ax = plt.subplots(ncols=3, figsize=(20,5))

fig.suptitle('Score Distributions of Students Based on Whether They Took the Course
or not')

a= sns.histplot(df, x='math score', ax=ax[0], hue='test preparation course', elemen
t='step')
b= sns.histplot(df, x='reading score', ax=ax[1], hue='test preparation course', ele
ment='step')
c= sns.histplot(df, x='writing score', ax=ax[2], hue='test preparation course', ele
ment='step')

sns.move_legend(a, "upper left", bbox_to_anchor=(0, 1))
sns.move_legend(b, "upper left", bbox_to_anchor=(0, 1))
sns.move_legend(c, "upper left", bbox_to_anchor=(0, 1))
```



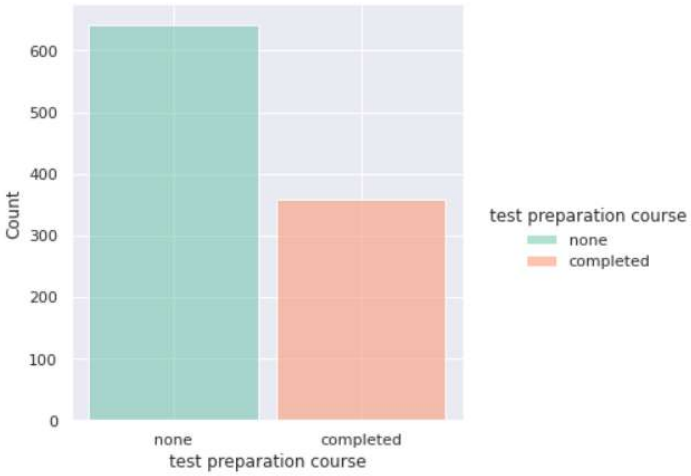

Если посмотреть на созданные нами гистограммы, то видно, что у студентов, посещавших курс (оранжевый цвет), наблюдается небольшой сдвиг вправо по сравнению со студентами, не посещавшими курс (зеленый цвет). Распределение зеленого цвета накладывается на распределение оранжевого цвета в частях с более низким баллом, а распределение оранжевого цвета затем накладывается на распределение зеленого цвета в частях с более высоким баллом, что означает, что если студент посещал курс, то, скорее всего, он получит более высокий балл, чем те студенты, которые его не посещали.

Тем не менее, есть много студентов, которые не посещали курс, и у них были более чем средние баллы, а некоторые даже сдали тесты на отлично. Чтобы

выяснить, верна ли наша гипотеза, мы должны визуализировать больше данных.

```
In [7]:
sns.displot(df, x='test preparation course', hue='test preparation course', shrink=
0.9)
```

```
Out[7]:
<seaborn.axisgrid.FacetGrid at 0x7fd0f1ed2910>
```




Поскольку категория "нет" имеет почти вдвое больше ответов, чем категория "завершено", это причина, по которой мы увидели, что распределение, окрашенное в зеленый цвет, было выше, чем распределение, окрашенное в оранжевый цвет.

## Пол

Следующая переменная, которую мы будем сравнивать, это пол, и посмотрим, есть ли связь между результатами тестов и полом.

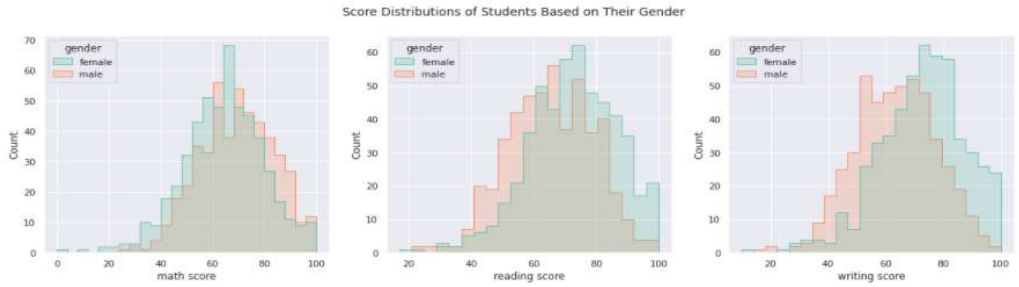
In [8]:

```
fig, ax = plt.subplots(ncols=3, figsize=(20,5))

fig.suptitle('Score Distributions of Students Based on Their Gender')

a= sns.histplot(df, x='math score', ax=ax[0], hue='gender', element='step')
b= sns.histplot(df, x='reading score', ax=ax[1], hue='gender', element='step')
c= sns.histplot(df, x='writing score', ax=ax[2], hue='gender', element='step')

sns.move_legend(a, "upper left", bbox_to_anchor=(0, 1))
sns.move_legend(b, "upper left", bbox_to_anchor=(0, 1))
sns.move_legend(c, "upper left", bbox_to_anchor=(0, 1))
```



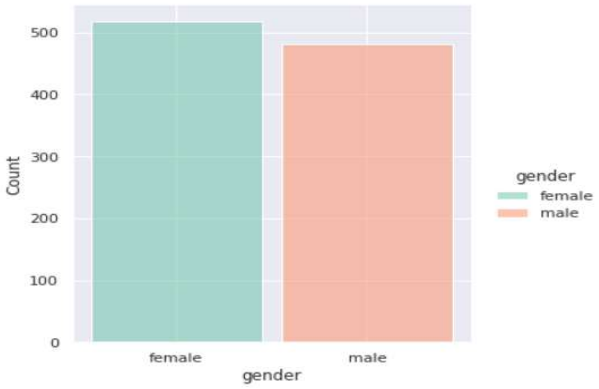

Исходя из того, что мы можем наблюдать в этих данных, студенты-мужчины, сдававшие тест по математике, получили немного более высокие баллы, чем студенты-женщины, хотя студенты-женщины, сдававшие тесты по чтению и письму, получили немного более высокие баллы, чем студенты-мужчины.

In [9]:

```
sns.displot(df, x='gender', hue='gender', shrink=0.9)
```

Out[9]:

```
<seaborn.axisgrid.FacetGrid at 0x7fd0f0355bd0>
```



Поэтому я также проверил распределение по полу студентов, и распределение кажется почти таким же.

## Раса/этническая принадлежность

Теперь мы добавим еще одну переменную для сравнения, которая называется "раса/этническая принадлежность" и относится к расе/этнической принадлежности студентов, которые делятся на 5 групп следующим образом:

Группа A

Группа B

Группа C

Группа D

Группа E

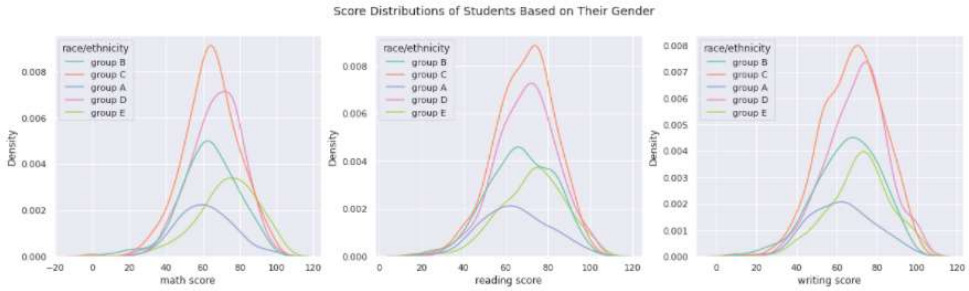
In [10]:

```
fig, ax = plt.subplots(ncols=3, figsize=(20,5))

fig.suptitle('Score Distributions of Students Based on Their Gender')

a= sns.kdeplot(data=df, x='math score', ax=ax[0], hue='race/ethnicity')
b= sns.kdeplot(data=df, x='reading score', ax=ax[1], hue='race/ethnicity')
c= sns.kdeplot(data=df, x='writing score', ax=ax[2], hue='race/ethnicity')

sns.move_legend(a, "upper left", bbox_to_anchor=(0, 1))
sns.move_legend(b, "upper left", bbox_to_anchor=(0, 1))
sns.move_legend(c, "upper left", bbox_to_anchor=(0, 1))
```



Наблюдая за графиком kde выше, можно заметить небольшой сдвиг между различными группами. Группа E имеет наибольший сдвиг вправо по сравнению с другими группами. Следовательно, если вы относитесь к группе E по расе/этнической принадлежности, у вас будет больше шансов получить более высокие баллы, чем у тех, кто к ней не относится.

## Уровень образования родителей

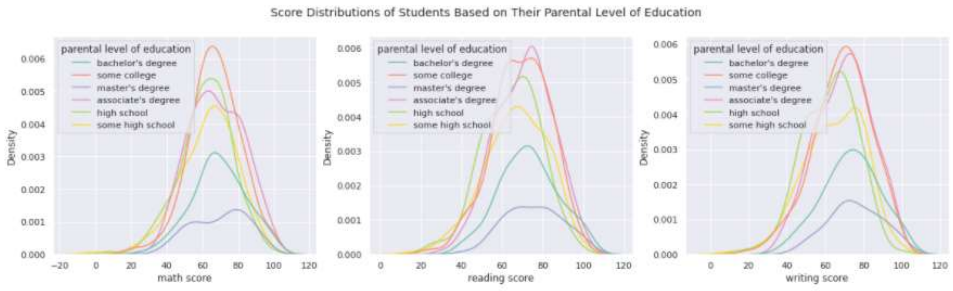
In [11]:

```
fig, ax = plt.subplots(ncols=3, figsize=(20,5))

fig.suptitle('Score Distributions of Students Based on Their Parental Level of Education')

a= sns.kdeplot(data=df, x='math score', ax=ax[0], hue='parental level of education')
b= sns.kdeplot(data=df, x='reading score', ax=ax[1], hue='parental level of education')
c= sns.kdeplot(data=df, x='writing score', ax=ax[2], hue='parental level of education')

sns.move_legend(a, "upper left", bbox_to_anchor=(0, 1))
sns.move_legend(b, "upper left", bbox_to_anchor=(0, 1))
sns.move_legend(c, "upper left", bbox_to_anchor=(0, 1))
```




Судя по приведенному выше графику, все они близки друг к другу, и это означает, что уровень образования родителей не так сильно влияет на результаты тестов.

## Обед

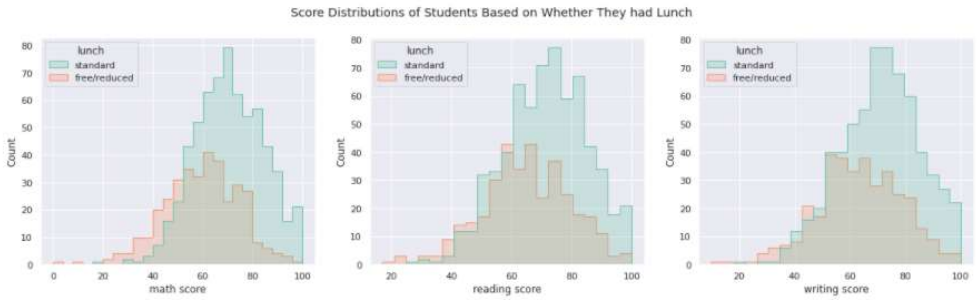
In [12]:

```
fig, ax = plt.subplots(ncols=3, figsize=(20,5))

fig.suptitle('Score Distributions of Students Based on Whether They had Lunch')

a= sns.histplot(df, x='math score', ax=ax[0], hue='lunch', element='step')
b= sns.histplot(df, x='reading score', ax=ax[1], hue='lunch', element='step')
c= sns.histplot(df, x='writing score', ax=ax[2], hue='lunch', element='step')

sns.move_legend(a, "upper left", bbox_to_anchor=(0, 1))
sns.move_legend(b, "upper left", bbox_to_anchor=(0, 1))
sns.move_legend(c, "upper left", bbox_to_anchor=(0, 1))
```




Совершенно очевидно, что студенты, которые стандартно обедали, получили более высокий балл, чем те, кто не обедал.

## 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