Кейс №1

РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

Валерия Ефимова

ЗАДАЧА

Провести анализ образовательных данных и выявить закономерности, влияющие на академическую успеваемость студентов.

ЭТАП 1: ВЫБОР ДАТАСЕТА

Для анализа была выбрана **тема образования.** С помощью сайта Kaggle были найдены несколько датасетов, которые потенциально могли бы быть интересными для рассмотрения.

```
[9] description_df1
```

	Признак	Тип данных	Шкала
0	form_no	int	Номинальная
1	name	string	Номинальная, 4343 уникальных значений
2	category	string	Номинальная
3	minority	string	Номинальная
4	gender	string	Номинальная, 2 уник. значения
5	hs_total	int	Относительная, 150-479
6	hs_pass_year	int	Порядковая, 2009-2018
7	first_choice_sub	string	Номинальная
8	first_choice_marks	int	Относительная
9	second_choice_sub	string	Номинальная

```
[11] description_df2
```

	Признак	Тип данных	Шкала
0	gender	string	Номинальная, 2 уник. значения
1	race/ethnicity	string	Номинальная
2	parental level of education	string	Порядковая
3	lunch	string	Номинальная, 2 уник. значения
4	test preparation course	string	Номинальная, 2 уник. значения
5	math score	int	Относительная, 0-100
6	reading score	int	Относительная, 17-100
7	writing score	int	Относительная, 10-100

```
description_df3
```

	Признак	Тип данных	Шкала
0	Marital status	int	Номинальная, 1-6
1	Application mode	int	Номинальная, 1-18
2	Application order	int	Порядковая, 1-9
3	Course	int	Номинальная
4	Daytime/evening attendance	int	Номинальная, 0-1
5	Previous qualification	int	Номинальная, 1-17
6	Nacionality	int	Номинальная, 1-21
7	Mother's qualification	int	Номинальная, 1-29
8	Father's qualification	int	Номинальная, 1-34
9	Mother's occupation	int	Номинальная, 1-32
10	Father's occupation	int	Номинальная, 1-46
11	Displaced	int	Номинальная, 0-1
12	Educational special needs	int	Номинальная, 0-1
13	Debtor	int	Номинальная, 0-1
14	Tuition fees up to date	int	Номинальная, 0-1
15	Gender	int	Номинальная, 0-1

Был взят датасет №2, так как он оказался наиболее удачным. Датасет №1 содержит много сомнительной информации, а датасет №3 слишком обработан, трудно понять смысл многих столбцов.

ЭТАП 1: ВЫБОР ДАТАСЕТА

Датасет №2 содержит информацию об учениках средней школы и их результатах по трём предметам: математика, чтение и письмо.

Особенности датасета:

– все числовые оценки находятся в шкале от 0 до 100 баллов.
– все признаки категориального типа представлены в текстовом формате и были закодированы при необходимости.
– данные не содержат пропусков и являются чистыми для анализа.

Датасет хорошо подходит для:
– анализа зависимости между подготовкой и результатами,
– исследования различий по полу,
– корреляционного анализа между предметами.

ЭТАП 2: ПРЕДОБРАБОТКА ДС

– Проверено наличие пропусков и выбросов.

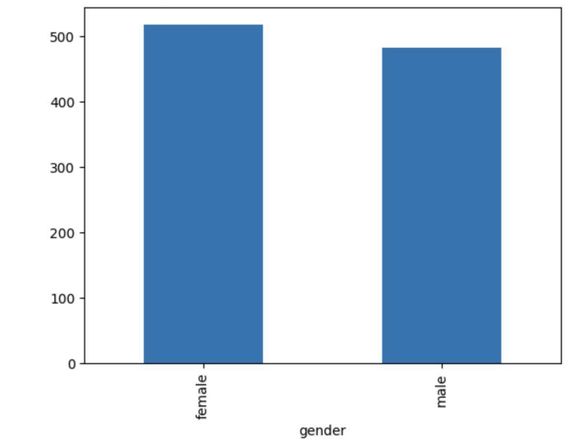
– Удалены нерелевантные признаки

– Проведена визуализация распределения оценок и сбалансированности признаков (например, по полу).

Проверим датасет на сбалансированность относительно признака 'gender'

```
df['gender'].value_counts().plot(kind='bar')
```

```
<Axes: xlabel='gender'>
```



Для анализа сформулированных гипотез нам не понадобятся некоторые столбцы, поэ

```
[16] df = df.drop(columns=['race/ethnicity', 'parental level of education', 'l
```

Проверим столбцы датасета на наличие нан-значений:

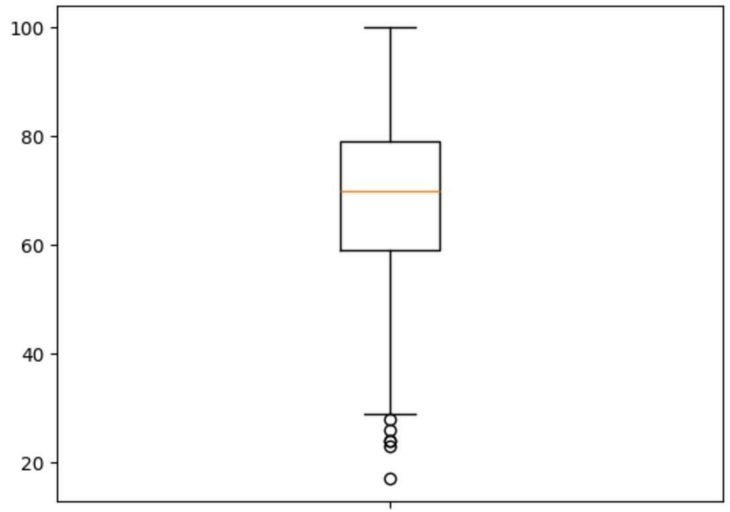
```
[17] df.isna().sum()
```

	0
gender	0
test preparation course	0
math score	0
reading score	0
writing score	0

dtype: int64

```
plt.boxplot(x=df['reading score'])
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x7bc64bad0a10>,
  <matplotlib.lines.Line2D at 0x7bc64bad1490>],
 'caps': [<matplotlib.lines.Line2D at 0x7bc64bad1e90>,
  <matplotlib.lines.Line2D at 0x7bc64bad2850>],
 'boxes': [<matplotlib.lines.Line2D at 0x7bc64bac0c90>],
 'medians': [<matplotlib.lines.Line2D at 0x7bc64bad3310>],
 'fliers': [<matplotlib.lines.Line2D at 0x7bc64bad3cd0>],
 'means': []}
```



ЭТАП 3: ФОРМУЛИРОВКА ГИПОТЕЗ

Гипотеза 1: Студенты, завершившие подготовительный курс, демонстрируют выше средние баллы по математике, чтению и письму, чем студенты, не проходившие этот курс.

Гипотеза 2: Баллы по математике, чтению и письму у студентов будут коррелировать между собой, то есть студенты, показавшие высокие результаты по одному предмету, склонны показать высокие результаты и по другим предметам.

Гипотеза 3: Студенты, не завершившие подготовительный курс, имеют большее распределение баллов в своих оценках, чем те, кто завершил курс.

Гипотеза 4: Мужчины в среднем получают более высокие баллы по математике, а женщины - по чтению и письму, поэтому средний балл женщин будет выше.

ЭТАП 4: ПРОВЕРКА ГИПОТЕЗ

В ходе проекта использовались следующие инструменты и методы анализа данных:

Инструменты:

– Python — основной язык анализа
– Pandas — для загрузки, очистки и обработки данных
– NumPy — для вычислений и статистики
– Matplotlib и Seaborn — для построения графиков, гистограмм и boxplot-диаграмм
– Scikit-learn (LabelEncoder) — для кодирования категориальных признаков
– Scipy (ttest_ind) — для проведения статистических тестов (t-тест)

Методы анализа:

– Описательная статистика (средние, медианы, размах)
– Визуализация распределений (гистограммы, boxplot)
– Корреляционный анализ (коэффициенты Пирсона и тепловые карты)
– Статистическая проверка гипотез с использованием t-критерия Стьюдента
– Группировка данных по категориям и сравнение значений между группами
– Обработка категориальных данных через кодирование признаков

ЭТАП 5: ВЫВОДЫ

1. Подготовка играет ключевую роль в повышении результатов.
2. Академические успехи по одному предмету часто сопровождаются успехами по другим.
3. Гендерные различия в результатах экзаменов имеют устойчивую статистическую основу.
4. Датасет показал полезность для педагогического и психологического анализа.