

**Федеральное государственное бюджетное образовательное
учреждение высшего образования
«РЫБИНСКИЙ ГОСУДАРСТВЕННЫЙ АВИАЦИОННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ имени П. А. СОЛОВЬЁВА»**

Институт информационных технологий и систем управления

Кафедра математического и программного обеспечения электронных
вычислительных средств

**Лабораторная работа №2
Обработка данных в датасете "Абитуриенты"**

по дисциплине
Методы и алгоритмы анализа данных

Выполнили студенты группы ИВМ-22

Перхуров В. А.
Беляев А. Е.

Проверил:

Кулиманов И. Е.

Рыбинск 2022

Задание

Постановка задачи:

Цели:

Провести анализ и обработку данных из заданного датасета "Абитуриенты"

Задачи:

1. Какая причина выбора школы была самой частой? В качестве ответа приведите соответствующее значение признака.
2. Найдите количество студентов, у родителей которых нет никакого образования.
3. Найдите минимальный возраст учащегося школы Mousinho da Silveira.
4. Найдите количество студентов, имеющих нечетное число пропусков.
5. Найдите разность между средними итоговыми оценками студентов, состоящих и не состоящих в романтических отношениях. В качестве ответа приведите число, округленное до двух значащих цифр после запятой.
6. Сколько занятий пропустило большинство студентов с самым частым значением наличия внеклассных активностей?
 - Определить самое частое значение наличия внеклассных активностей (да или нет).
 - Для группы студентов, соответствующей этому значению, рассмотреть значения признака «число пропусков».
 - Для каждого значения числа пропусков посчитать, сколько студентов ему соответствует.
 - Выбрать значение числа пропусков с наибольшим числом студентов.

1 Описание датасета "Абитуриенты"

Набор данных представлен в CSV-файле. Каждая строка наборов данных содержит следующие поля:

1. school - тип школы ("GP Gabriel Pereira или "MS Mousinho da Silveira)
2. sex - пол ("F female или "M male)
3. age - возраст (от 15 до 22)
4. address - откуда студент ("U urban или "R rural)
5. famsize - размер семьи ("LE3 меньше или равно 3 или "GT3 больше 3)
6. Pstatus - в каких отношениях родители ("T живут вместе "A раздельно)
7. Medu - образование матери (0 - никакого, 1 - начальное образование (4 класса), 2 - от 5 до 9 классов, 3 - среднеспециальное или 4 - высшее)
8. Fedu - образование отца (0 - никакого, 1 - начальное образование (4 класса), 2 - от 5 до 9 классов, 3 - среднеспециальное или 4 - высшее)
9. Mjob - работа матери ("teacher "health"care related, civil "services"(e.g. administrative or police), "at_home"or "other")
10. Fjob - работа отца ("teacher "health"care related, civil "services"(e.g. administrative or police), "at_home"or "other")
11. reason - причина выбора школы (близко к дому — "home репутация школы — "reputation предпочтение некоторым предметам - "course"или "other")
12. guardian - опекун ("mother "father"или "other")
13. traveltime - время от дома до школы (1 - меньше 15 мин., 2 - 15 до 30 мин., 3 - 30 мин. до 1 часа, или 4 - больше 1 часа)
14. studytime - количество часов обучения в неделю (1 - меньше 2 часов, 2 - от 2 до 5 часов, 3 - от 5 до 10 часов, или 4 - больше 10 часов)
15. failures - количество ранее не сданных предметов (n if $1 \leq n < 3$, else 4)
16. schoolsup - дополнительные занятия (yes or no)
17. famsup - помощь от семьи при выполнении заданий (yes or no)
18. paid - дополнительные платные занятия (yes or no)

19. activities - внеклассная деятельность (yes or no)
20. nursery - посещал детский сад (yes or no)
21. higher - желание высшего образования (yes or no)
22. internet - домашний интернет (yes or no)
23. romantic - состоит в романтических отношениях (yes or no)
24. famrel - насколько хороши отношения в семье (от 1 - очень плохие до 5 - превосходные)
25. freetime - наличие свободного времени после школы (от 1 - очень мало до 5 - очень много)
26. goout - гуляет с друзьями (от 1 - редко до 5 - очень часто)
27. Dalc - употребление алкоголя в будние дни (от 1 - очень редко до 5 - очень часто)
28. Walc - употребление алкоголя в выходные (от 1 - очень редко до 5 - очень часто)
29. health - текущее состояние здоровья (от 1 - очень плохое до 5 - очень хорошее)
30. absences - количество школьных пропусков (от 0 до 93)
31. G1 - оценка за первый семестр (от 0 до 20)
32. G2 - оценка за второй семестр (от 0 до 20)
33. G3 - итоговая оценка (от 0 до 20)

2 Описание выполнения работы

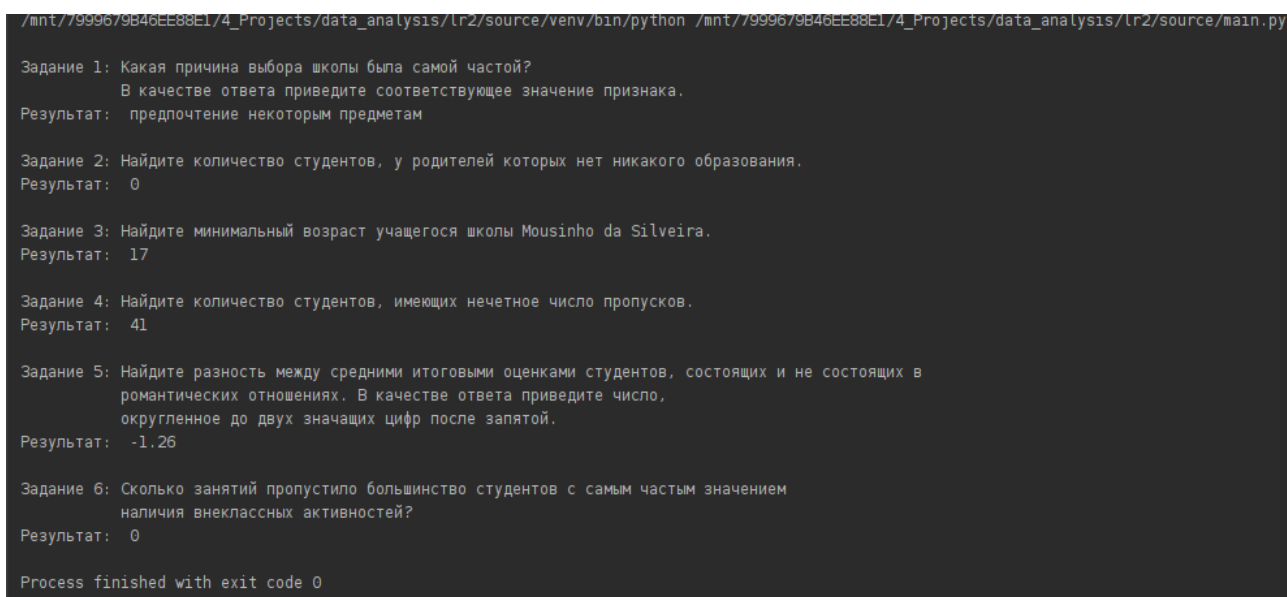
2.1 Реализация программы для анализа датасета "Абитуриенты"

2.1.1 Используемые библиотеки

При решении данной задачи была использована библиотека pandas (анализ данных).

2.1.2 Ход работы и результат выполнения программы

В ходе выполнения работы была написана программа, которая выводит результат анализа датасета в лог. На рисунке 1 показан данный вывод.



```
/mnt/7999679B46EE88E1/4_Projects/data_analysis/lr2/source/venv/bin/python /mnt/7999679B46EE88E1/4_Projects/data_analysis/lr2/source/main.py

Задание 1: Какая причина выбора школы была самой частой?
В качестве ответа приведите соответствующее значение признака.
Результат: предпочтение некоторым предметам

Задание 2: Найдите количество студентов, у родителей которых нет никакого образования.
Результат: 0

Задание 3: Найдите минимальный возраст учащегося школы Mousinho da Silveira.
Результат: 17

Задание 4: Найдите количество студентов, имеющих нечетное число пропусков.
Результат: 41

Задание 5: Найдите разность между средними итоговыми оценками студентов, состоящих и не состоящих в романтических отношениях. В качестве ответа приведите число, округленное до двух значащих цифр после запятой.
Результат: -1.26

Задание 6: Сколько занятий пропустило большинство студентов с самым частым значением наличия внеклассных активностей?
Результат: 0

Process finished with exit code 0
```

Рис. 1: Результат работы программы

2.1.3 Исходный текст программы

```
import pandas as pd # анализ данных

# Найти наиболее распространённое значение в датасете.
# Принимает набор данных и название столбца для определения
# его самого частотного значения.
def FindMostCommonValues(data, column_name):
    return pd.value_counts(data[column_name].values, sort=True).axes[0][0]
```

```

# Задание 1:
#         Какая причина выбора школы была самой частой?
#         В качестве ответа приведите соответствующее значение признака.
def FindMostCommonReasonForChoosingSchool(main_df):
    reason_values = {
        "home": "близко к дому",
        "reputation": "репутация школы",
        "course": "предпочтение некоторым предметам",
        "other": "предпочтение некоторым предметам",
    }

    # определение самой частой причины выбора школы
    reason = FindMostCommonValues(main_df, 'reason')

    print("""
Задание 1: Какая причина выбора школы была самой частой?
           В качестве ответа приведите соответствующее значение признака.
Результат: """, reason_values[reason])

# Задание 2:
#         Найдите количество студентов, у родителей которых
#         нет никакого образования.
def FindNumberOfStudentsWhoseParentsHaveNoEducation(main_df):
    no_education_value = 0
    data = main_df[(main_df['Medu'] == no_education_value) &
                    (main_df['Fedu'] == no_education_value)]

    print("""
Задание 2: Найдите количество студентов, у родителей которых
           нет никакого образования.
Результат: """, len(data))

# Задание 3:
#         Найдите минимальный возраст учащегося школы
#         Mousinho da Silveira.
def FindMinimumAgeOfStudentAtMousinhoDaSilveiraSchool(main_df):
    school_name = 'MS'
    data = main_df[(main_df['school'] == school_name)]

    print("""
Задание 3: Найдите минимальный возраст учащегося школы

```

Mousinho da Silveira.

Результат: "", min(data['age']))

Задание 4:

Найдите количество студентов,
имеющих нечетное число пропусков.

```
def FindNumberOfStudentsWhoHaveAnOddNumberOfAbsences(main_df):  
    data = main_df[main_df['absences'] % 2 != 0]
```

```
    print("""
```

Задание 4: Найдите количество студентов, имеющих нечетное число пропусков.

Результат: "", len(data))

Задание 5:

Найдите разность между средними итоговыми оценками студентов,
состоящих и не состоящих в романтических отношениях.
В качестве ответа приведите число,
округленное до двух значащих цифр после запятой.

```
def FindDifferenceBetweenAverageFinalGradesOfStudentsInAndOutOfRomanticRel  
    data = main_df.groupby('romantic').describe()  
    # "{:.2f}".format - округление до 2-х символов после запятой  
    result_difference = "{:.2f}".format(data['G3', 'mean']['yes'] -  
                                         data['G3', 'mean']['no'])
```

```
    print("""
```

Задание 5: Найдите разность между средними итоговыми оценками студентов,
состоящих и не состоящих в романтических отношениях.

В качестве ответа приведите число,
округленное до двух значащих цифр после запятой.

Результат: "", result_difference)

Задание 6:

Сколько занятий пропустило большинство студентов
с самым частым значением наличия внеклассных активностей?

```
def HowManyClassesDidMostStudentsWithMostFrequentValueOfHavingExtracurricu  
    activities_is_exist = FindMostCommonValues(main_df, 'activities')  
    data_with_activities = main_df[main_df['activities'] ==  
                                     activities_is_exist]  
    number_of_students_by_absences =  
        pd.value_counts(data_with_activities['absences'].values,  
                        sort=True)
```

```

    print("""
Задание 6: Сколько занятий пропустило большинство студентов
           с самым частым значением наличия внеклассных активностей?
Результат: """, number_of_students_by_absences.axes[0][0])

# Основная функция
def main():
    # 0 - Читаем датасет
    df = pd.read_csv('./math_students.csv',
                     escapechar='\'',
                     low_memory=False)

    # Задание 1
    FindMostCommonReasonForChoosingSchool(df)
    # Задание 2
    FindNumberOfStudentsWhoseParentsHaveNoEducation(df)
    # Задание 3
    FindMinimumAgeOfStudentAtMousinhoDaSilveiraSchool(df)
    # Задание 4
    FindNumberOfStudentsWhoHaveAnOddNumberOfAbsences(df)
    # Задание 5
    FindDifferenceBetweenAverageFinalGradesOfStudentsInAndOutOfRomanticRel
    # Задание 6
    HowManyClassesDidMostStudentsWithMostFrequentValueOfHavingExtracurricu

# Запускаем программу
main()

```


3 Вывод.

В результате выполнения лабораторной работы были проведёны анализ и обработка данных из заданного датасета "Абитуриенты".

В ходе выполнения работы была написана программа, которая выполняет чтение датасета из файла в формате ".csv" и реализуют заданный функционал:

1. Найти самую частую причину выбора школы.
2. Найти количество студентов, у которых родители не имеют никакого образования.
3. Найти минимальный возраст учащегося школы Mousinho da Silveira.
4. Найти количество студентов, имеющих нечетное число пропусков.
5. Найти разность между средними итоговыми оценками студентов, состоящих и не состоящих в романтических отношениях.
6. Найти количество занятий, пропущенных большинством студентов, у которых есть (или нет) внеклассная активность (в зависимости от того, что чаще встречается).