

**Федеральное государственное автономное образовательное
учреждение высшего образования “Национальный
исследовательский университет
“Высшая школа экономики”
Московский институт электроники и математики им. А.Н. Тихонова
НИУ ВШЭ
Департамент компьютерной инженерии**

Курс: «Проектный семинар «Python в науке о данных»

**Руководство пользователя
по проекту “Political Coordinates”**

Группа:

БИВ225

Номер бригады:

Директор бригады:

Носов Иван Александрович

+7 (924) 301-27-51

ianosov@edu.hse.ru

Состав бригады:

Носов Иван Александрович,

Пантякова Дарья Евгеньевна,

Лифановский Дмитрий Валентинович

Руководитель:

Полякова Марина Васильевна

МОСКВА 2023

Содержание

1. Назначение программы
2. Технические требования
3. Описание каталогов
4. Описание структуры базы данных
5. Установка и запуск приложения
 - 5.1 Подготовка к запуску
 - 5.2 Главный экран
 - 5.3 Работа с базой данных
 - 5.4 Статистический отчет
 - 5.5 Сводная таблица
 - 5.6 Кластеризованная столбчатая диаграмма
 - 5.7 Категоризированная гистограмма
 - 5.8 Категоризированная диаграмма Бокса-Вискера
 - 5.9 Категоризированная диаграмма рассеивания
 - 5.10 Встроенное прохождение теста
 - 5.11 Настройки программы

1. Назначение программы

Приложение "Political Coordinates" является специализированным информационно-аналитическим инструментом для профессионалов, который предназначен для сбора, обработки и хранения данных, полученных от респондентов или данных, существующих в базе данных. Оно также включает функциональность для анализа данных в виде графических отчетов, таких как таблицы, диаграммы и гистограммы. После анализа отчетов, данные могут сравниваться между собой для получения дополнительных выводов.

2. Технические требования

Для корректной работы программы у пользователя / оператора¹ должен быть в наличии персональный компьютер с минимальными характеристиками:

- процессор x86 с тактовой частотой, не менее 1 ГГц;
- оперативная память объемом, не менее 1 Гб;
- монитор, мышь, клавиатура;
- ОС MS Windows 10.

Приоритетным требованием является наличие операционной системы Windows, в которой имеется возможность установки интерпретатора Python 3 или дистрибутива Anaconda версии, актуальной для даты начала работы проектного семинара.

Получить данный интерпретатор и проверить совместимость своей операционной системы с ним можно на официальном сайте разработчика интерпретатора по ссылке: <https://www.python.org/downloads/>. Для Anaconda актуальной является версия "Anaconda3-2022.05".

3. Описание каталогов

Так как программа распространяется и предоставляется пользователю / оператору в виде архива, необходимо знать его структуру. Пользователю доступен архив со следующей структурой каталогов:

- Work - основной каталог;
 - Data - содержит базу данных;
 - Library - содержит библиотеку стандартных функций;
 - Notes - содержит документацию;
 - Script - содержит специализированный модуль и файл с определением параметров настройки приложения.

4. Описание структуры базы данных

Вся работа программы осуществляется при помощи анализа и модификаций записей в базе данных. При этом база данных, аналогично пункту

¹ в данном контексте эти понятия являются эквивалентными

3, также имеет определенную структуру. И для эффективной работы с программой необходимо знать структуру базы данных. База данных содержит 8 основных полей, а именно:

1. id - индивидуальный номер каждой записи;
2. gender - содержит информацию о гендере респондента;
3. field - научное направление, которое глобально характеризует ОП респондента;
4. university - учебное заведение респондента;
5. course - курс обучения респондента;
6. x - содержит координату по оси X, которая была ранее вычислена определенным способом и характеризует политические предпочтения респондента;
7. y - аналогично столбцу "x", только по оси Y;
8. z - аналогично столбцу "x", только по оси Z.

5. Установка и запуск приложения

5.1 Подготовка к запуску

Если пользователь установил дистрибутив Anaconda, то большинство необходимых для работы программы библиотек, такие как NumPy, pandas, matplotlib, SciPy и многие другие уже установлены. Если же пользователь установил только интерпретатор Python3, то ему необходимо дополнительно установить нужные библиотеки. Сделать это можно при помощи менеджера пакетов pip, который в большинстве компьютеров под управлением ОС Windows уже установлен, и терминала.

Сначала необходимо запустить терминала и в необходимой директории ввести команду "pip install <имя библиотеки>". Найти имена и версии всех необходимых библиотек можно в загруженном архиве в файле "requirements.txt".

Проверить наличие библиотек и их версии можно также в терминале, используя команду "pip freeze" в нужной директории.

Список необходимых библиотек и их версий из файла "requirements.txt":

- numpy 1.24.3
- pandas 2.0.1
- python-dateutil 2.8.2
- pytz 2023.3
- six 1.16.0
- tzdata 2023.3

После всех проделанных действий, описанных выше, пользователю необходимо запустить приложение из командной строки командой "python main.py". При этом пользователю необходимо находиться в нужной директории, где и располагается установочный файл. Его можно найти через путь

Загрузки\political_coord\work\). Но в большинстве случаев его расположение зависит от того, как пользователь распаковал полученный архив.

5.2 Главный экран

После запуска приложения из командной строки командой "python main.py" на мониторе перед пользователем появится стартовый экран приложения и первая страница взаимодействия с данными по совместительству (рис. 1).

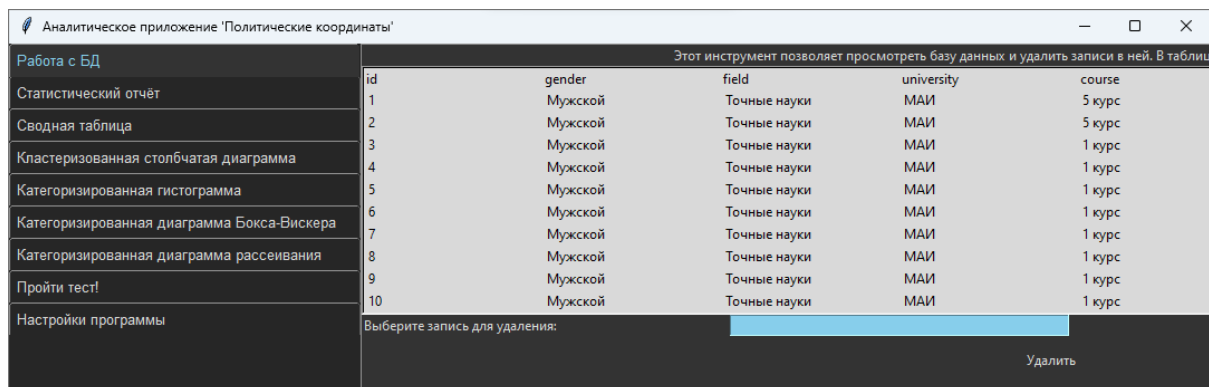


Рис. 1 Главный экран и простой текстовый отчет

Можно заметить, что главный экран разделен на 2 части: список всех доступных видов текстовых, графических отчетов, других разделов и рабочее пространство, в котором можно взаимодействовать с этими отчетами.

5.3 Работа с базой данных

Продолжая рассматривать первую начальную страницу можно обратить внимание на модификацию записей в базе данных, а именно на операцию проекции и сокращения таблицы. Данная операция производится путем удаления столбцов или строк, поэтому перед пользователем предстает сама база данных в виде таблицы и текстовое поле, в которое можно ввести номер, т.е. id, записи и при нажатии на кнопку "Удалить" произвести удаление соответствующей записи (рис. 2). Данное текстовое поле на вход принимает только целые положительные числа, которые и являются номерами записями, при попытке ввести данные другого типа никаких операций не произойдет и база данных останется в прежнем виде.

Остальные модификации записей в базе данных, их добавление и т.п. будут рассматриваться в других разделах, так как конкретно этот раздел имеет своей целью демонстрацию текстового отчета в виде таблице, полученной только путем вычеркивания части строк / столбцов из базы данных.

id	gender	field	university	course
1	Мужской	Точные науки	МАИ	5 курс
2	Мужской	Точные науки	МАИ	1 курс
3	Мужской	Точные науки	МАИ	1 курс
4	Мужской	Точные науки	МАИ	1 курс
5	Мужской	Точные науки	МАИ	1 курс
6	Мужской	Точные науки	МАИ	1 курс
7	Мужской	Точные науки	МАИ	1 курс
8	Мужской	Точные науки	МАИ	1 курс
9	Мужской	Точные науки	МАИ	1 курс
10	Мужской	Точные науки	МАИ	1 курс

Выберите запись для удаления:

Удалить

id	gender	field	university	course
1	Мужской	Точные науки	МАИ	1 курс
2	Мужской	Точные науки	МАИ	1 курс
3	Мужской	Точные науки	МАИ	1 курс
4	Мужской	Точные науки	МАИ	1 курс
5	Мужской	Точные науки	МАИ	1 курс
6	Мужской	Точные науки	МАИ	1 курс
7	Мужской	Точные науки	МАИ	1 курс
8	Мужской	Точные науки	МАИ	1 курс
9	Мужской	Точные науки	МАИ	1 курс
10	Мужской	Точные науки	МАИ	5 курс

Выберите запись для удаления:

1

Удалить

Рис. 2 Демонстрация удаления записи №1 в базе данных

5.4 Статистический отчет

Далее во многих отчетах будут появляться качественные и количественные атрибуты, т.е. атрибуты, которые нельзя измерить или выразить числом, а которые могут быть только описаны качественно или атрибуты, которые можно измерить и выразить числом соответственно.

В данном разделе приложение составляет статистический отчет по каждому полю базы данных. Перед пользователем на экране появляется два комбобокса или выпадающих списка для каждого типа атрибутов, т.е. отдельно для качественных и количественных, из которых можно выбрать название поля и составить по нему статистический отчет.

Для качественной переменной создается отчет в виде таблицы (рис. 3), первый столбец которой содержит значения переменной, второй - частоты, а третий — процент количества указанных объектов от их общего числа.

Качественный атрибут:	field
Количественный атрибут:	z
Создать отчет для качественного атрибута	Создать отчет для количественного атрибута

Значение	Частоты	Процент
Точные науки	526	46.964285714285715
Социальные науки	241	21.517857142857146
Гуманитарные науки	234	20.892857142857142
Естественные науки	119	10.625

Рис. 3 Статистический отчет для качественных переменных

Для количественных же переменных пользователь выбирает необходимое поле (значение координат по осям x, y или z) и нажимает на кнопку "Создать отчет для количественных атрибутов". После этого появляется таблица (рис. 4) с двумя колонками: названием статистической переменной и ее значением. В таблице имеются такие переменные как общее количество данного атрибута, его минимальное и максимальное значение, арифметическое среднее, выборочная дисперсия и стандартное отклонение.

Количественный атрибут:	y
Создать отчет для качественного атрибута	Создать отчет для количественного атрибута

Статистика	Значение
Всего	1094.0
Среднее	-0.3587751371115174
Отклонение	3.7163163345605086
Минимальное	-10.0
25%	-3.0
50%	-0.5
75%	2.0
Максимальное	10.0

Рис. 4 Демонстрация работы статистического отчета для количественных атрибутов

5.5 Сводная таблица

В данном разделе перед пользователем предстает 4 выпадающих списка, а именно: ось, первый атрибут, второй атрибут и метод агрегации. Данный анализ производится для любой пары качественных переменных, содержащихся в столбцах gender, field, university, course. Также помимо выбора пары качественных атрибутов, выбирается и одна из осей, которая и определяет политические координаты.

Далее следует выбор метода агрегации, т.е. метода обработки данных, при котором множество значений, относящихся к нескольким объектам,

объединяются в одно значение, которое представляет собой общую характеристику группы объектов.

Объединение в одно значение может происходить через сумму параметров, их минимум или максимум, а также через среднее значение и медиану.

В итоге пользователю предоставляется отчет в виде таблицы (рис. 5) с полями, которые относятся к паре выбранных атрибутов, относительно параметра, который характеризуется значением координат по выбранной оси и методом агрегации.

Ось:

Первый атрибут:

Второй атрибут:

Метод агрегации:

Создать сводную таблицу

field	Женский	Мужской
Гуманитарные науки	-204.0	-45.5
Естественные науки	-47.0	-30.5
Социальные науки	-202.0	-79.5
Точные науки	-80.5	296.5

Рис. 5 Результат анализа данных через "сводную таблицу"

5.6 Кластеризованная столбчатая диаграмма

В данном разделе пользователю необходимо выбрать два качественных атрибута в двух выпадающих списках сверху, относительно которых и будет строиться кластеризованная столбчатая диаграмма. В виде качественных атрибутов принимаются поля gender, field, university, course. В итоге получается график (рис. 6), который позволяет сравнивать значения нескольких групп или категорий данных. Он имеет несколько столбцов, каждый из которых соответствует отдельной категории, а на оси Y отображаются значения этих категорий.

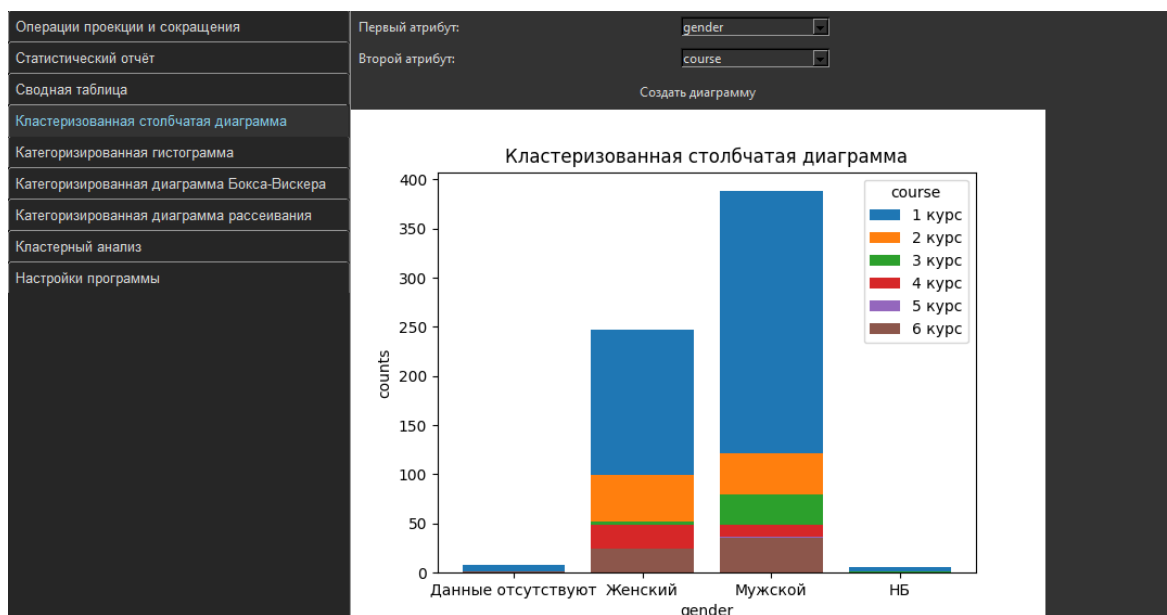


Рис. 6 Кластеризованная столбчатая диаграмма

5.7 Категоризированная гистограмма

По началу данный способ анализа данных выглядит схожим с предыдущим, только в этом отчете происходит сравнение количественных атрибутов с качественными. После выбора нужных атрибутов в выпадающих списках вверху рабочего пространства и нажатия на кнопку создания отчета перед пользователем предстает гистограмма (рис. 7), которая позволяет отображать распределение значений на определенном диапазоне. Она имеет несколько столбцов, каждый из которых соответствует определенной категории, а на оси X отображаются значения этих категорий.

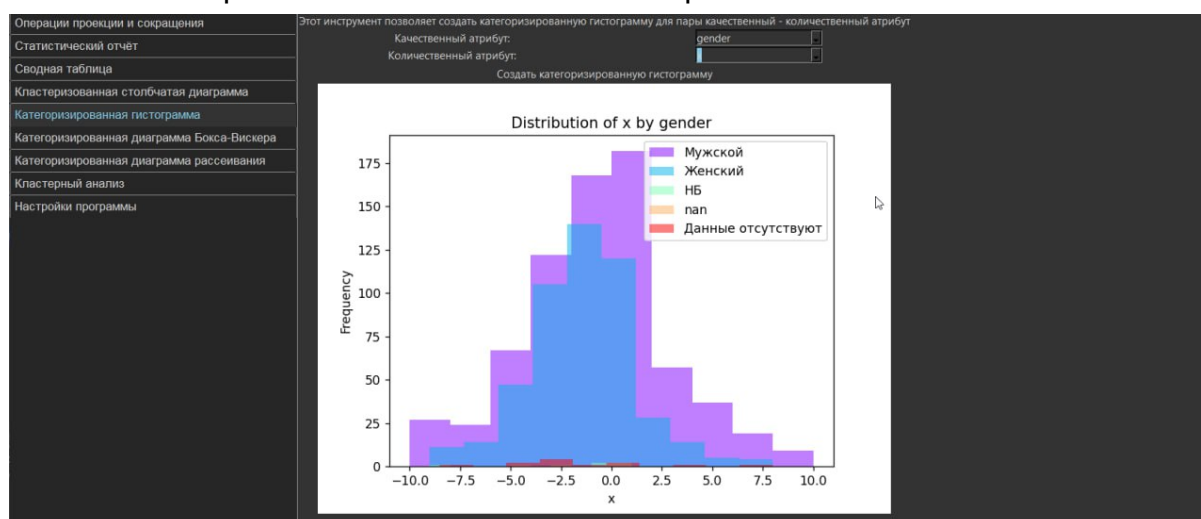


Рис. 7 Категоризированная гистограмма

5.8 Категоризированная диаграмма Бокса-Вискера

При создании данного типа отчета, пользователь проделывает те же действия, что и в предыдущем разделе, так как здесь также имеются комбобоксы

с качественным и количественным атрибутом и кнопкой создания диаграммы. Отличается только метод построения диаграммы.

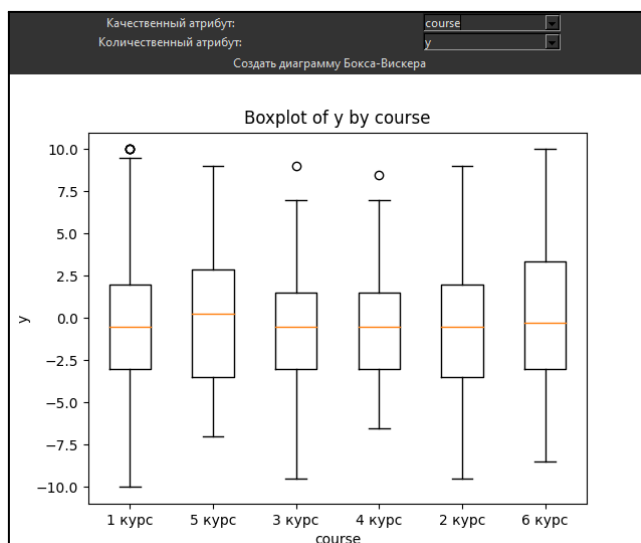


Рис. 8 Пример работы отчета "Категоризированная диаграмма Бокса-Вискера"

На графике (рис. 8) построены боксы (box plots), каждый из которых представляет собой диапазон, в котором находится большинство данных. Горизонтальная линия внутри бокса - это медиана (середина) распределения. Отметки за боксом - это выбросы, то есть необычные значения в данных.

5.9 Категоризированная диаграмма рассеивания

При открытии этого окна вверху экрана предстает 3 комбобокса: качественный атрибут, количественный атрибут 1, количественный атрибут 2. После выбора необходимых пользователю атрибутов и нажатия на кнопку "Создать диаграмму рассеивания" на экране строится график (рис. 9), который позволяет визуализировать соотношение между двумя количественными переменными и одной категориальной переменной. Эта диаграмма также называется точечной диаграммой с группировкой, так как данные точки на графике группируются по категориальной переменной.

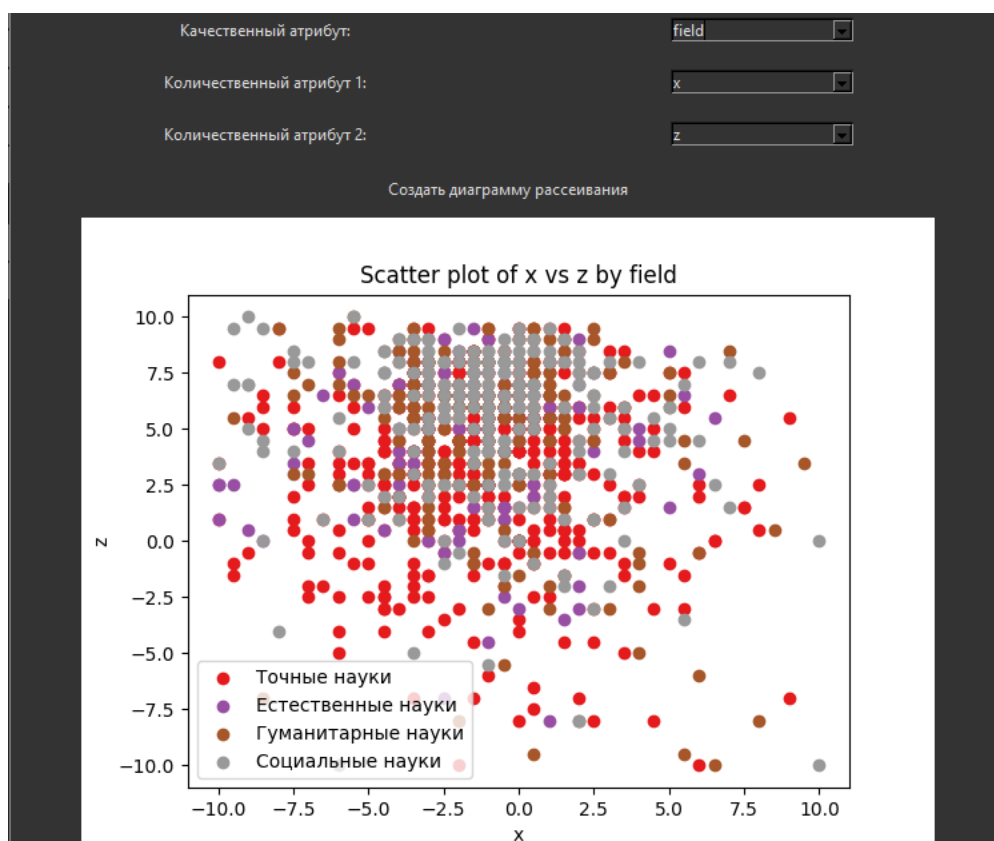


Рис. 9 Категоризированная диаграмма рассеивания

5.10 Встроенное прохождение теста

При открытии данного раздела пользователю предоставляется возможность пройти тест с последующим занесением результатов в базу данных. Сверху имеются 4 выпадающих списка с вводом информации о респонденте, а именно его пол, направление обучения, ВУЗ и курс обучения (рис. 10). Далее пользователю демонстрируется вопрос и 4 варианта ответа на него с возможностью выбора одного ответа. После предоставления ответа на вопрос необходимо нажать на кнопку "Следующий вопрос" и повторить эти действие несколько раз, пока не закончатся все вопросы. Как пользователь ответил на все вопросы, ему демонстрируется его результат в виде значений по трем координатам (рис. 11). Результаты также заносятся в базу данных.

Женский

Точные науки

НИЯУ МИФИ

4 курс

Повышение налогов для богатых людей несёт вред обществу.

☐ Полностью согласен

☐ Скорее согласен

☒ Скорее не согласен

☐ Категорически не согласен

Следующий вопрос

Рис. 10 Прохождение теста

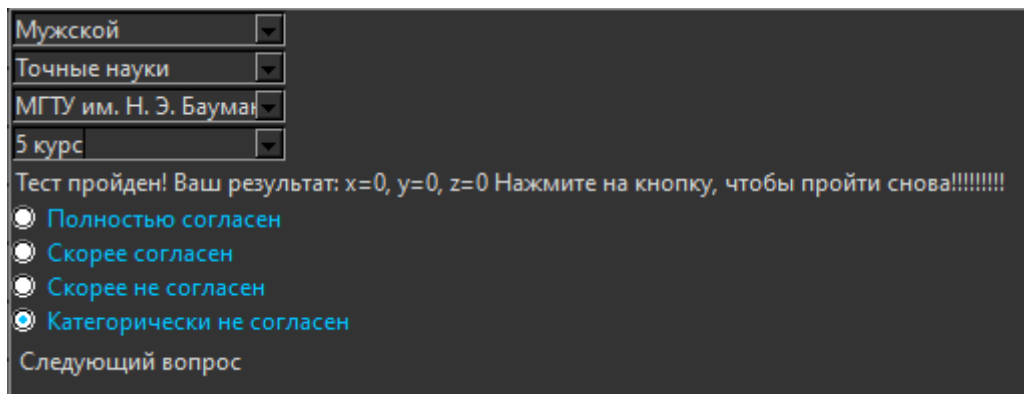


Рис. 11 Демонстрация результата

5.11 Настройки программы

В последнем разделе производится настройка программы, а именно базы данных. Изначально приложение работает с заранее загруженной базой данных, располагающейся в ...work/data/data.csv. Но при необходимости можно загрузить и свою базу данных, которая будет иметь такие же поля, как и исходная.

Для этого в представленном текстовом поле (рис. 12) необходимо указать адрес расположения заранее созданной вами базы данных. Далее уже в другом текстовом поле нужно дать название этому кейсу, чтобы потом постоянно не указывать путь к нужной базе данных. Особых правил по даче названий не существует. После этого нужно нажать на кнопку "Сохранить настройки" для их сохранения. Далее уже можно продолжать работы с выбранной вами базой данных.

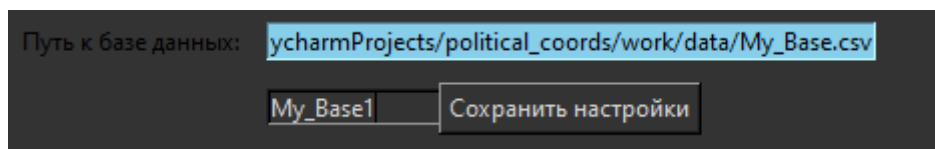


Рис. 12 Пример создания настройки для баз данных