Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение

высшего образования

«Пермский национальный исследовательский политехнический университет»

Кафедра информационных технологий и автоматизированных систем

**ОТЧЁТ ПО САМОСТОЯТЕЛЬНОЙ РАБОТЕ № 3**

|  |
| --- |
| по дисциплине «Технологии анализа данных DataMining»  Тема: «Использование библиотеки numpy» |
|  |

|  |  |
| --- | --- |
|  | Выполнил студент гр. АСУ8-24-1м  Пельц Данил Андреевич  (Фамилия И.О.)  (номер зачетной книжки)  Проверил доцент каф. ИТАС  (должность)  Городилов Алексей Юрьевич  (Фамилия И.О)  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  (оценка)  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  (дата, подпись) |

Пермь 2025 г.

Цель работы: Исследовать и проанализировать данные о производстве, потреблении и зависимости прибыли от скидок с помощью методов обработки массивов и аппроксимации, реализуя вычисления и визуализацию с использованием библиотек numpy, scipy и matplotlib.

Задачи:

1. Загрузить и обработать данные о производстве и потреблении электроэнергии различных стран, а также данные о прибыли в зависимости от скидок.
2. Рассчитать средние показатели производства и потребления электроэнергии за определённый период и определить ключевые статистические характеристики по разным критериям.
3. Построить и решить системы линейных уравнений для нахождения коэффициентов квадратичных и кубических полиномов, аппроксимирующих зависимость прибыли от скидок.
4. Оценить качество полученных моделей с помощью меры квадратичного отклонения (RSS), визуализировать результаты и сделать прогнозы прибыли для заданных значений скидок.

**Решение задания 1:**

1. **Загрузка данных**

Для загрузки данных из файлов была применена функция np.genfromtxt. Также были использованы параметры для разделения по столбцам и правильной обработки типов данных (рис. 1):



Рис. 1

1. **Среднее ежегодное производство и потребление за последние 5 лет**

Было выччеслено среднее значение по последним 5 годам с помощью функции np.nanmean с параметром axis=1 (по строкам) (рис. 2):

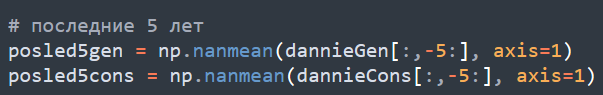


Рис. 2

**3.1 Суммарное потребление электроэнергии за каждый год**

Сумма значений по всем странам для каждого года с помощью np.nansum по оси 0 (рис. 3):



Рис. 3

**3.2 Максимальное количество электроэнергии, произведённое одной страной за один год**

Был найден максимум с игнорированием NaN через np.nanmax (рис. 4):

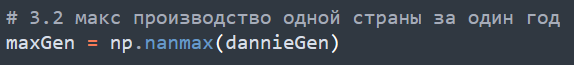


Рис. 4

**3.3 Страны с производством более 500 млрд кВт\*ч в среднем за последние 5 лет (рис. 5)**

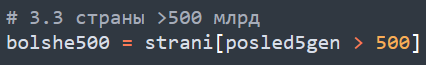


Рис. 5

**3.4 10% стран с наибольшим средним потреблением за последние 5 лет**

Был вычислен 90-й процентиль и фильтрация страны по массиву средних потреблений (рис. 6):

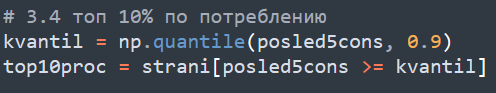


Рис. 6

**3.5 Страны, которые увеличили производство в 2021 году более чем в 10 раз по сравнению с 1992 (рис. 7)**

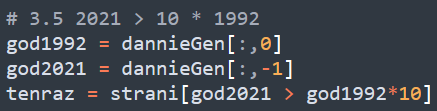


Рис. 7

**3.6 Страны, потратившие более 100 млрд кВт\*ч и при этом произведшие меньше, чем потратили (рис. 8)**

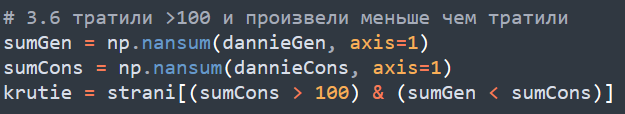


Рис. 8

**3.7 Страна с максимальным потреблением в 2020 году (рис. 9)**

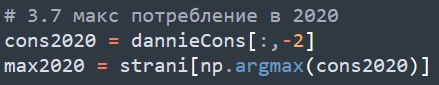


Рис. 9

**4. Вывод значений и визуализация суммарного потребления по годам (рис. 10)**

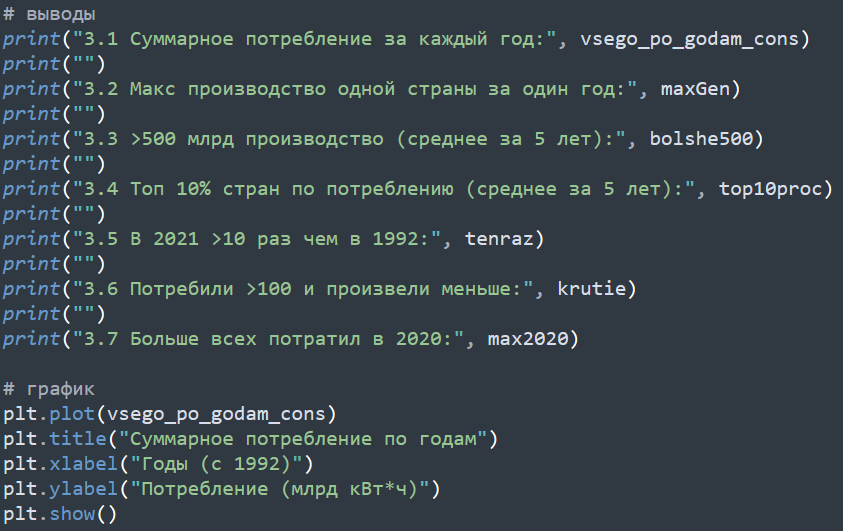


Рис. 10

**Решение задания 2:**

1. **Загрузка и подготовка данных**

Были импортированы данные из файла «data2.csv» с помощью функции np.genfromtxt. Для корректной обработки строк с символами BOM и названиями месяцев реализована очистка и преобразование значений скидок в числовой формат. Это позволило получить массивы скидок и прибыли, пригодные для дальнейшего анализа и апроксимации (рис. 1):

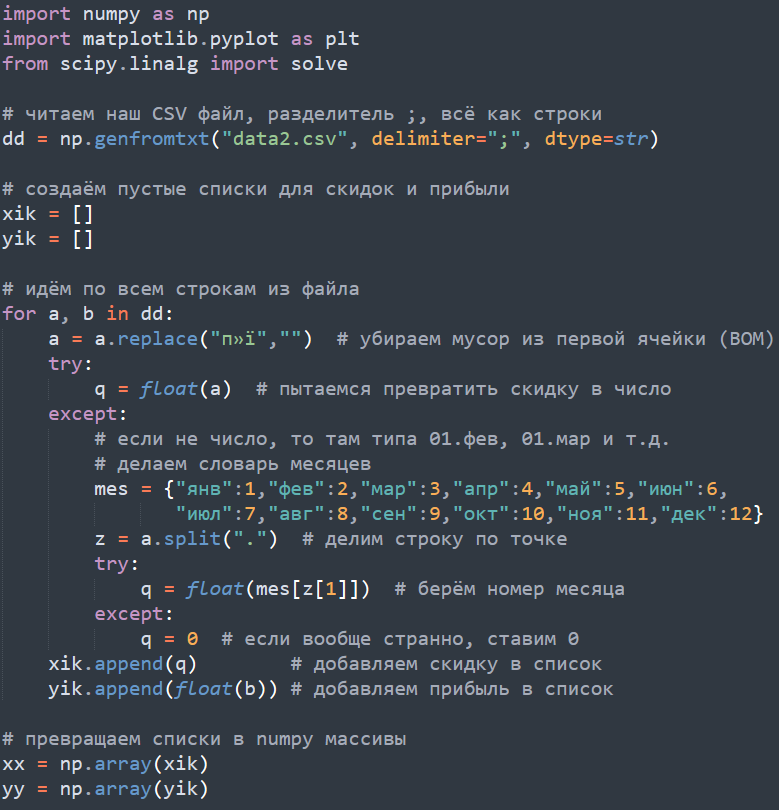


Рис. 1

1. **Формирование системы уравнений для квадратичного полинома**

Для обеспечения точности модели выбрал три точки, равномерно распределённые по диапазону данных (начало, середина, конец). Построена система линейных уравнений, в которой значения полинома в этих точках должны совпадать с исходными данными. Такой выбор обеспечивает более представительное аппроксимирование зависимости (Рис. 2):

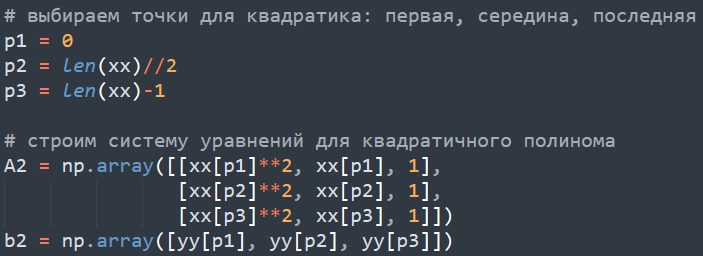


Рис. 2

1. **Решение СЛУ для нахождения коэффициентов квадратичного полинома**

Решение системы при помощи scipy.linalg.solve позволило получить коэффициенты a2,a1,a0a2,a1,a0 квадратичного полинома. Это даёт аналитическую формулу, описывающую приближённую зависимость прибыли от скидки (рис. 3):

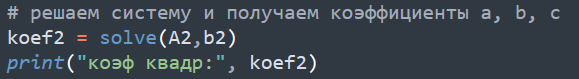


Рис. 3

1. **Вычисление значений квадратичного полинома на всех точках и оценка ошибки**

Полученный полином применён ко всем исходным значениям скидок, что позволило сравнить прогнозируемые и реальные прибыли. Подсчитано квадратичное отклонение (RSS), что служит мерой качества аппроксимации — чем меньше RSS, тем лучше модель описывает данные (рис. 4):

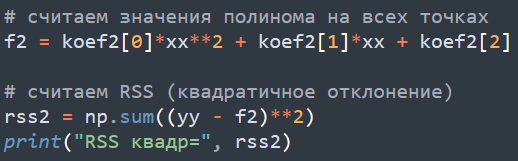


Рис. 4

1. **Формирование системы уравнений для кубического полинома**

Для улучшения точности модели добавлена четвертая точка, что соответствует полиному третьей степени. Аналогично сформирована и решена система линейных уравнений, что позволяет получить более гибкую формулу зависимости (рис. 5):

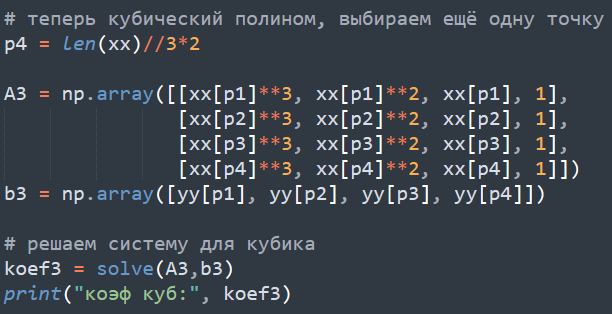


Рис. 5

1. **Вычисление значений кубического полинома и оценка ошибки**

Расчёт значений кубического полинома позволил провести сравнение с реальными данными и вычислить RSS для оценки точности модели. Это даёт основание оценить, насколько сложная модель лучше или хуже отражает зависимость (рис. 6):

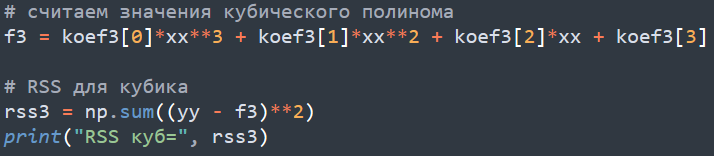


Рис. 6

1. **Визуализация исходных данных и обеих аппроксимаций**

Построены графики, позволяющие наглядно сравнить исходные данные с моделями. Визуальное представление помогает понять, какая из моделей лучше подходит под данные и выявить особенности аппроксимации (рис. 7):



Рис. 7

1. **Выбор лучшей модели по минимальному RSS и прогноз прибыли при новых скидках**

Сравнение RSS двух моделей позволило выбрать наиболее точную. На её основе рассчитаны прогнозируемые значения прибыли при скидках 6% и 8%. Это демонстрирует практическое применение модели для бизнес-прогнозов (рис. 8):



Рис. 8

**Вывод:**

В лабораторной работе я разобрался, как загрузить и обработать данные по электроэнергии и прибыли из разных файлов. Считал средние значения за последние годы и нашёл интересные показатели, вроде стран с большим производством или высоким потреблением. Затем построил простые модели — квадратичную и кубическую — чтобы понять, как прибыль зависит от скидок. Для этого решил систему уравнений, чтобы получить формулы, которые приблизительно описывают данные.

Потом проверил, какая модель лучше, считая ошибку (RSS), и сравнил графики. Оказалось, что одна из моделей подходит точнее, и на её основе сделал прогнозы прибыли при новых скидках. В итоге эта работа показала, что даже простые математические приёмы с помощью кода помогают получить полезные выводы из цифр и понять, как меняются данные.