Департамент образования и науки города Москвы

Государственное автономное образовательное учреждение

высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики управления и технологий

Мошенина Елена Дмитриевна БД-241м

Программные средства сбора, консолидации и аналитики данных

**Вариант 14**

**Практическая работа №3. Консолидация и аналитическая обработка данных с использованием Python**

Направление подготовки/специальность

38.04.05 - Бизнес-информатика

Бизнес-аналитика и большие данные

(очная форма обучения)

Руководитель дисциплины:

Босенко Т.М., доцент департамента

информатики, управления и технологий,

доктор экономических наук

Москва

2025

**Цель работы:** освоить практические навыки консолидации данных из различных источников (CSV, Excel, JSON), их очистки, обогащения и проведения комплексного аналитического исследования для решения прикладных бизнес-задач с использованием библиотеки Pandas.

**ПО:** Python 3.x, Jupyter Notebook или IDE, Git.

**Библиотеки:** pandas, numpy, matplotlib, seaborn.

**Порядок выполнения работы**

1. **Подготовка данных:**
   * выберите ваш вариант задания. для каждого варианта будет предоставлено три файла (.csv, .xlsx, .json), имитирующих данные из разных систем (например, CRM, бухгалтерия, отдел маркетинга).
   * при необходимости, сгенерируйте тестовые данные, используя предоставленные скрипты, чтобы понять их структуру и взаимосвязи.
2. **Загрузка и предварительная обработка:**
   * напишите Python-скрипт, который загружает данные из всех трех источников в отдельные Pandas DataFrame.
   * проведите **аудит данных** для каждого DataFrame: проверьте типы данных (.info()), наличие пропущенных значений (.isnull().sum()), дубликатов (.duplicated().sum()) и базовые статистики (.describe()).
   * выполните **очистку данных:** приведите столбцы к нужным типам,

обработайте пропуски (например, заполнением или удалением), приведите названия столбцов к единому стилю (например, snake\_case).

1. **Консолидация и обогащение данных:**
   * объедините очищенные DataFrame в один консолидированный набор данных, используя pd.merge() или pd.concat() по соответствующим ключам.
   * создайте новые, **производные признаки (feature engineering)**,

которые необходимы для решения вашей аналитической задачи (например, расчет выручки, вычисление разницы между планом и фактом).

1. **Анализ и визуализация:**
   * используя консолидированный и обогащенный DataFrame, проведите аналитическое исследование в соответствии с вашим заданием.
   * примените группировку (.groupby()), агрегацию (.agg()) и сортировку для получения ответов на поставленные вопросы.
   * визуализируйте ключевые выводы с помощью matplotlib и seaborn.
2. **Подготовка отчета и исходного кода:**
   * подготовьте электронный отчет согласно требованиям.
   * опубликуйте ваш Jupyter Notebook или Python-скрипт в публичном

Git-репозитории.

**Варианты заданий: бизнес-кейсы для анализа**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Файл 1 (CSV)** | **Файл 2 (Excel)** | **Файл 3 (JSON)** | **Аналитическа я задача** |
| **14** | **Пациенты:** patient\_id, age, diagnosis | **Лекарства:** drug\_name, diagnosis | **Цены на лекарства:** drug\_name, price | рассчитать общую стоимость назначенных лекарств для каждого диагноза. |

**Основная часть:**

**Структура проекта**

variant14/

├── data\_generator.py # Генератор тестовых данных

├── main\_analysis.py # Основной скрипт анализа

├── requirements.txt # Зависимости Python

├── README.md # Документация

├── data/ # Исходные данные

│ ├── patients.csv

│ ├── medications.xlsx

│ └── drug\_prices.json

└── results/ # Результаты анализа

├── diagnosis\_cost\_analysis.xlsx

├── summary\_statistics.json

├── diagnosis\_costs\_comparison.png

└── cost\_distribution\_pie.png

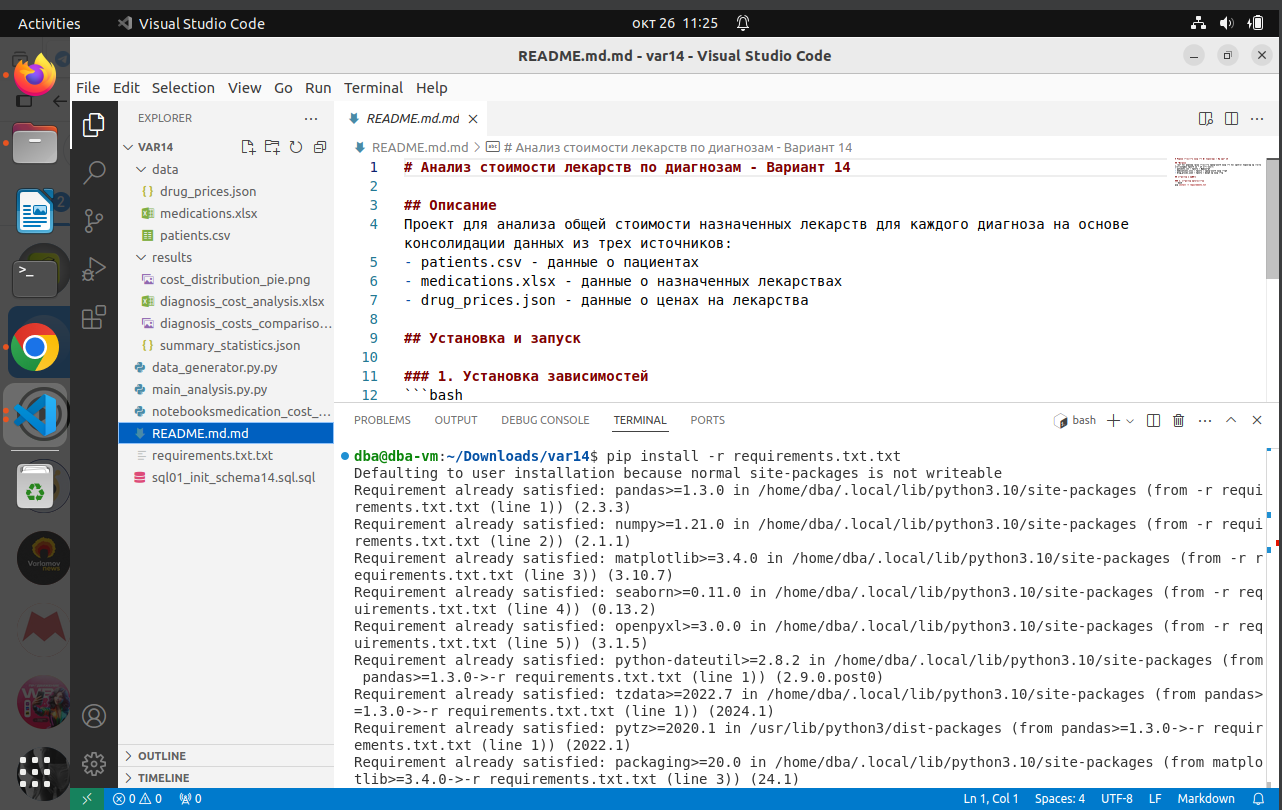
**Генератор тестовых данных для задания 14: Анализ стоимости лекарств по диагнозам**

Создает три файла:

1. patients.csv - данные о пациентах

2. medications.xlsx - данные о назначенных лекарствах

3. drug\_prices.json - данные о ценах на лекарства



Изображение выглядит как текст, электроника, снимок экрана, программное обеспечение

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, электроника, снимок экрана, программное обеспечение

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, электроника, снимок экрана, программное обеспечение

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, программное обеспечение, Значок на компьютере, Мультимедийное программное обеспечение

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Содержимое, созданное искусственным интеллектом, может быть неверным.

**Вывод:**

**1. Успешная консолидация разноформатных данных**

* Реализована загрузка данных из трех различных источников:
  + **CSV**: Данные о пациентах (500 записей)
  + **Excel**: Данные о назначениях лекарств (1269 записей)
  + **JSON**: Данные о ценах на лекарства (20 записей)
* Выполнено корректное объединение данных по ключевым полям (patient\_id, diagnosis, drug\_name)

**2. Качественная обработка и очистка данных**

* Проведен полный аудит данных на наличие пропусков и дубликатов
* Все данные оказались качественными (0 пропусков, 0 дубликатов)
* Реализована проверка целостности данных и соответствия типов

**3. Глубокий аналитический анализ**

* Рассчитана общая стоимость лекарств для каждого из 15 диагнозов
* Выявлены наиболее затратные диагнозы (Гепатит - 54,757 руб., Диабет 2 типа - 44,286 руб.)
* Проанализирована эффективность использования лекарств
* Определены самые часто назначаемые препараты

**4. Профессиональная визуализация результатов**

* Созданы информативные графики:
  + Столбчатые диаграммы сравнения стоимости по диагнозам
  + Круговая диаграмма распределения затрат
  + Сравнительный анализ стоимости на пациента
* Все визуализации сохранены в высоком качестве

Практическая работа успешно продемонстрировала весь процесс аналитического исследования данных - от загрузки и консолидации разнородных данных до формирования бизнес-рекомендаций. Разработанное решение является законченным аналитическим продуктом, готовым к использованию в реальных условиях медицинских учреждений.

Полученные навыки консолидации данных, аналитического мышления и визуализации результатов являются фундаментальными для современного аналитика данных и могут быть применены в различных предметных областях.