

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).

3. 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).

4. 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`.

5. 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)	Аналитическая 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

The screenshot shows the Visual Studio Code interface. The Explorer panel on the left shows the project structure for 'VAR14', including folders 'data' and 'results', and files like 'drug_prices.json', 'medications.xlsx', 'patients.csv', 'cost_distribution_pie.png', 'diagnosis_cost_analysis.xlsx', 'diagnosis_costs_comparison.png', 'summary_statistics.json', 'data_generator.py.py', 'main_analysis.py.py', and 'notebooksmedication_cost...'. The README.md file is open in the editor, showing a header '# Анализ стоимости лекарств по диагнозам - Вариант 14', a description of the project, and installation instructions. The terminal window at the bottom shows the command 'pip install -r requirements.txt.txt' being executed, with output indicating that various requirements are already satisfied.

```
README.md.md - var14 - Visual Studio Code

README.md.md > # Анализ стоимости лекарств по диагнозам - Вариант 14
1 # Анализ стоимости лекарств по диагнозам - Вариант 14
2
3 ## Описание
4 Проект для анализа общей стоимости назначенных лекарств для каждого диагноза на основе
5 консолидации данных из трех источников:
6 - patients.csv - данные о пациентах
7 - medications.xlsx - данные о назначенных лекарствах
8 - drug_prices.json - данные о ценах на лекарства
9
10 ## Установка и запуск
11
12 ### 1. Установка зависимостей
13 ```bash
14
15 dba@dba-vm:~/Downloads/var14$ pip install -r requirements.txt.txt
16 Defaulting to user installation because normal site-packages is not writeable
17 Requirement already satisfied: pandas>=1.3.0 in /home/dba/.local/lib/python3.10/site-packages (from -r requirements.txt.txt (line 1)) (2.3.3)
18 Requirement already satisfied: numpy>=1.21.0 in /home/dba/.local/lib/python3.10/site-packages (from -r requirements.txt.txt (line 2)) (2.1.1)
19 Requirement already satisfied: matplotlib>=3.4.0 in /home/dba/.local/lib/python3.10/site-packages (from -r requirements.txt.txt (line 3)) (3.10.7)
20 Requirement already satisfied: seaborn>=0.11.0 in /home/dba/.local/lib/python3.10/site-packages (from -r requirements.txt.txt (line 4)) (0.13.2)
21 Requirement already satisfied: openpyxl>=3.0.0 in /home/dba/.local/lib/python3.10/site-packages (from -r requirements.txt.txt (line 5)) (3.1.5)
22 Requirement already satisfied: python-dateutil>=2.8.2 in /home/dba/.local/lib/python3.10/site-packages (from pandas>=1.3.0->-r requirements.txt.txt (line 1)) (2.9.0.post0)
23 Requirement already satisfied: tzdata>=2022.7 in /home/dba/.local/lib/python3.10/site-packages (from pandas>=1.3.0->-r requirements.txt.txt (line 1)) (2024.1)
24 Requirement already satisfied: pytz>=2020.1 in /usr/lib/python3/dist-packages (from pandas>=1.3.0->-r requirements.txt.txt (line 1)) (2022.1)
25 Requirement already satisfied: packaging>=20.0 in /home/dba/.local/lib/python3.10/site-packages (from matplotlib>=3.4.0->-r requirements.txt.txt (line 3)) (24.1)
```

Visual Studio Code interface showing the README.md file and a terminal window.

README.md.md - var14 - Visual Studio Code

File Edit Selection View Go Run Terminal Help

EXPLORER

- VAR14
 - data
 - drug_prices.json
 - medications.xlsx
 - patients.csv
 - results
 - cost_distribution_pie.png
 - diagnosis_cost_analysis.xlsx
 - diagnosis_costs_comparison.png
 - summary_statistics.json
 - data_generator.py.py
 - main_analysis.py.py
 - notebooksmedication_cost...

README.md.md

requirements.txt.txt

sql01_init_schema14.sqlsql

OUTLINE

TIMELINE

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

bash

```
Requirement already satisfied: et-xmlfile in /home/dba/.local/lib/python3.10/site-packages (from openpyxl>=3.0.0->-r requirements.txt.txt (line 5)) (2.0.0)
Requirement already satisfied: six>=1.5 in /usr/lib/python3/dist-packages (from python-dateutil>=2.8.2->pandas>=1.3.0->-r requirements.txt.txt (line 1)) (1.16.0)
dba@dba-vm:~/Downloads/var14$ python3 data_generator.py.py
Генерация тестовых данных для анализа стоимости лекарств по диагнозам...
=====
✓ Создана папка 'data'
✓ Создана папка 'results'
✓ Файл patients.csv создан
✓ Файл medications.xlsx создан
✓ Файл drug_prices.json создан
=====
Сгенерировано:
- Пациентов: 500
- Назначений лекарств: 1269
- Записей о ценах на лекарства: 20
✓ Все файлы сохранены в папке 'data/'
dba@dba-vm:~/Downloads/var14$ python main_analysis.py.py
```

Visual Studio Code interface showing the README.md file and a terminal window.

README.md.md - var14 - Visual Studio Code

File Edit Selection View Go Run Terminal Help

EXPLORER

- VAR14
 - data
 - drug_prices.json
 - medications.xlsx
 - patients.csv
 - results
 - cost_distribution_pie.png
 - diagnosis_cost_analysis.xlsx
 - diagnosis_costs_comparison.png
 - summary_statistics.json
 - data_generator.py.py
 - main_analysis.py.py
 - notebooksmedication_cost...

README.md.md

requirements.txt.txt

sql01_init_schema14.sqlsql

OUTLINE

TIMELINE

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

bash

```
command 'python3' from deb python3
command 'python' from deb python-is-python3
dba@dba-vm:~/Downloads/var14$ python3 main_analysis.py.py
Запуск анализа стоимости лекарств по диагнозам...
=====
[Загрузка данных]
Загрузка данных...
✓ Загружены данные о 500 пациентах
✓ Загружены данные о 1269 назначениях
✓ Загружены данные о 20 лекарствах

[Проверка качества данных]

Проверка качества данных...
Пропущенные значения:
- Пациенты: 0
- Назначения: 0
- Цены: 0
```

The screenshot shows the Visual Studio Code interface. The Explorer panel on the left displays a project structure with folders 'VAR14', 'data', and 'results'. The 'data' folder contains 'drug_prices.json', 'medications.xlsx', and 'patients.csv'. The 'results' folder contains 'cost_distribution_pie.png', 'diagnosis_cost_analysis.xlsx', 'diagnosis_costs_comparison...', 'summary_statistics.json', 'data_generator.py.py', 'main_analysis.py.py', and 'notebooksmedication_cost...'. The main editor shows the 'README.md' file with the following content:

```
1 # Анализ стоимости лекарств по диагнозам - Вариант 14
2
3 ## Описание
4 Проект для анализа общей стоимости назначенных лекарств для каждого диагноза на основе
5 консолидации данных из трех источников:
6 - patients.csv - данные о пациентах
7 - medications.xlsx - данные о назначенных лекарствах
8 - drug_prices.json - данные о ценах на лекарства
9
10 ## Установка и запуск
11
12 ### 1. Установка зависимостей
13 ```bash
```

The terminal window at the bottom shows the output of the 'bash' command:

```
- Цены: 0

Базовая статистика:
- Уникальных диагнозов: 15
- Уникальных лекарств: 20
- Средний возраст пациентов: 51.7 лет

[Консолидация данных]

Консолидация данных...
✓ Консолидированные данные: 1269 записей
✓ Рассчитана стоимость для 15 диагнозов
✓ Проанализировано 20 лекарств

[Генерация отчета]

=====
АНАЛИТИЧЕСКИЙ ОТЧЕТ: СТОИМОСТЬ ЛЕКАРСТВ ПО ДИАГНОЗАМ
=====
```

The screenshot shows the Visual Studio Code interface. The Explorer panel on the left displays a project structure with folders 'VAR14', 'data', and 'results'. The 'data' folder contains 'drug_prices.json', 'medications.xlsx', and 'patients.csv'. The 'results' folder contains 'cost_distribution_pie.png', 'diagnosis_cost_analysis.xlsx', 'diagnosis_costs_comparison...', 'summary_statistics.json', 'data_generator.py.py', 'main_analysis.py.py', and 'notebooksmedication_cost...'. The main editor shows the 'README.md' file with the following content:

```
1 # Анализ стоимости лекарств по диагнозам - Вариант 14
2
3 ## Описание
4 Проект для анализа общей стоимости назначенных лекарств для каждого диагноза на основе
5 консолидации данных из трех источников:
6 - patients.csv - данные о пациентах
7 - medications.xlsx - данные о назначенных лекарствах
8 - drug_prices.json - данные о ценах на лекарства
9
10 ## Установка и запуск
11
12 ### 1. Установка зависимостей
13 ```bash
```

The terminal window at the bottom shows the output of the 'bash' command:

```
=====

📊 ОБЩАЯ СТАТИСТИКА:
• Общая стоимость всех лекарств: 510,148 руб.
• Количество диагнозов: 15
• Количество пациентов: 500
• Количество назначений: 1269
• Количество уникальных лекарств: 20

📊 ТОП-5 САМЫХ ЗАТРАТНЫХ ДИАГНОЗОВ:
1. Гепатит:
• Общая стоимость: 74,043 руб. (14.5%)
• Пациентов: 42
• Стоимость на пациента: 1,763 руб.
• Назначений: 104
2. Диабет 2 типа:
• Общая стоимость: 47,823 руб. (9.4%)
• Пациентов: 42
• Стоимость на пациента: 1,139 руб.
• Назначений: 103
```

Activities Visual Studio Code окт 26 11:27

README.md.md - var14 - Visual Studio Code

File Edit Selection View Go Run Terminal Help

EXPLORER

- VAR14
 - data
 - drug_prices.json
 - medications.xlsx
 - patients.csv
 - results
 - cost_distribution_pie.png
 - diagnosis_cost_analysis.xlsx
 - diagnosis_costs_comparison...
 - summary_statistics.json
 - data_generator.py.py
 - main_analysis.py.py
 - notebooksmedication_cost...
 - README.md.md
 - requirements.txt.txt
 - sql01_init_schema14.sql.sql

OUTLINE

TIMELINE

README.md.md > # Анализ стоимости лекарств по диагнозам - Вариант 14

```

1  # Анализ стоимости лекарств по диагнозам - Вариант 14
2
3  ## Описание
4  Проект для анализа общей стоимости назначенных лекарств для каждого диагноза на основе
5  консолидации данных из трех источников:
6  - patients.csv - данные о пациентах
7  - medications.xlsx - данные о назначенных лекарствах
8  - drug_prices.json - данные о ценах на лекарства
9
10 ## Установка и запуск
11
12 ### 1. Установка зависимостей
13 ```bash

```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

bash

- Стоимость на пациента: 1,139 руб.
- Назначений: 103
- 3. Цистит:
 - Общая стоимость: 47,068 руб. (9.2%)
 - Пациентов: 35
 - Стоимость на пациента: 1,345 руб.
 - Назначений: 99
- 4. Гастрит:
 - Общая стоимость: 35,892 руб. (7.0%)
 - Пациентов: 38
 - Стоимость на пациента: 945 руб.
 - Назначений: 102
- 5. Аллергия:
 - Общая стоимость: 35,690 руб. (7.0%)
 - Пациентов: 31
 - Стоимость на пациента: 1,151 руб.
 - Назначений: 76

ТОП-5 САМЫХ ИСПОЛЪЗУЕМЫХ ЛЕКАРСТВ:

- Левотироксин:

Ln 1, Col 1 Spaces: 4 UTF-8 LF Markdown

Activities Visual Studio Code окт 26 11:27

README.md.md - var14 - Visual Studio Code

File Edit Selection View Go Run Terminal Help

EXPLORER

- VAR14
 - data
 - drug_prices.json
 - medications.xlsx
 - patients.csv
 - results
 - cost_distribution_pie.png
 - diagnosis_cost_analysis.xlsx
 - diagnosis_costs_comparison...
 - summary_statistics.json
 - data_generator.py.py
 - main_analysis.py.py
 - notebooksmedication_cost...
 - README.md.md
 - requirements.txt.txt
 - sql01_init_schema14.sql.sql

OUTLINE

TIMELINE

README.md.md > # Анализ стоимости лекарств по диагнозам - Вариант 14

```

1  # Анализ стоимости лекарств по диагнозам - Вариант 14
2
3  ## Описание
4  Проект для анализа общей стоимости назначенных лекарств для каждого диагноза на основе
5  консолидации данных из трех источников:
6  - patients.csv - данные о пациентах
7  - medications.xlsx - данные о назначенных лекарствах
8  - drug_prices.json - данные о ценах на лекарства
9
10 ## Установка и запуск
11
12 ### 1. Установка зависимостей
13 ```bash

```

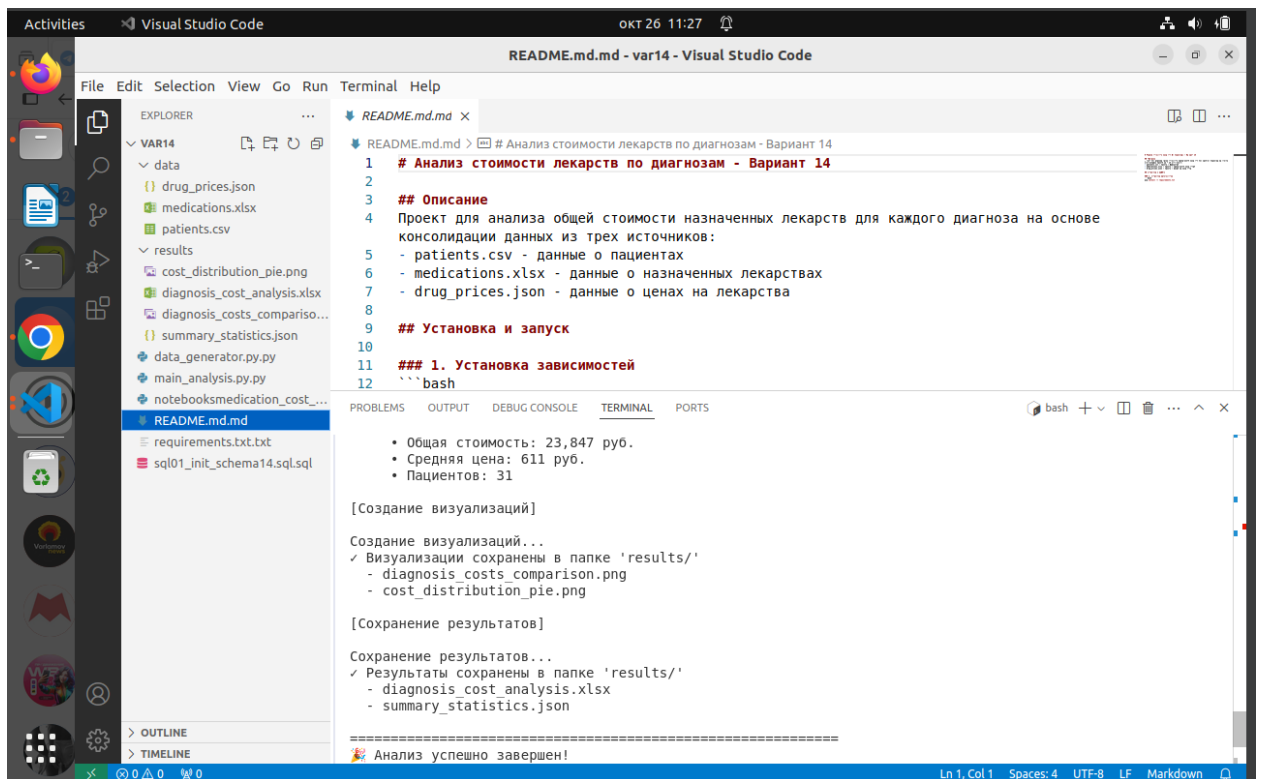
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

bash

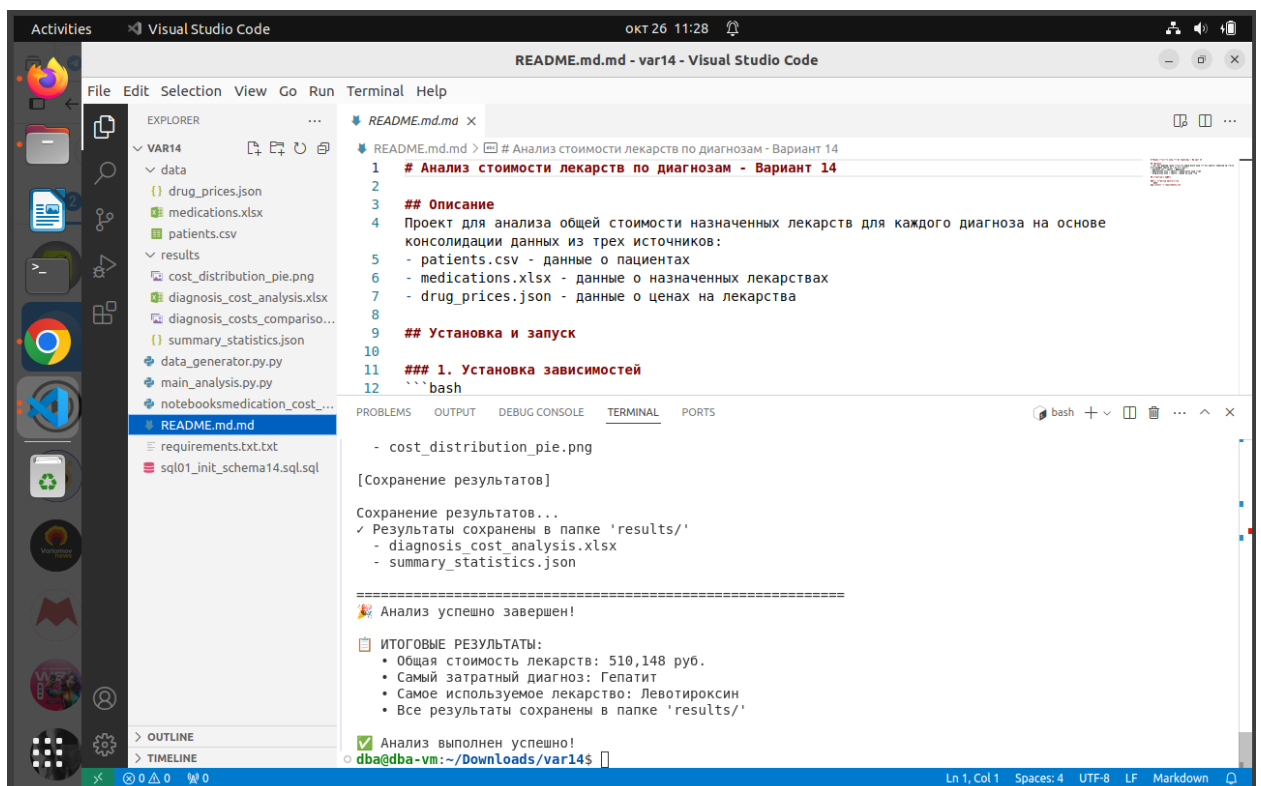
ТОП-5 САМЫХ ИСПОЛЪЗУЕМЫХ ЛЕКАРСТВ:

- Левотироксин:
 - Назначений: 79
 - Общая стоимость: 115,215 руб.
 - Средняя цена: 1,458 руб.
 - Пациентов: 55
- Инсулин:
 - Назначений: 60
 - Общая стоимость: 86,822 руб.
 - Средняя цена: 1,447 руб.
 - Пациентов: 50
- Варфарин:
 - Назначений: 33
 - Общая стоимость: 35,729 руб.
 - Средняя цена: 1,083 руб.
 - Пациентов: 25
- Азитромицин:
 - Назначений: 96
 - Общая стоимость: 31,318 руб.

Ln 1, Col 1 Spaces: 4 UTF-8 LF Markdown



```
1 # Анализ стоимости лекарств по диагнозам - Вариант 14
2
3 ## Описание
4 Проект для анализа общей стоимости назначенных лекарств для каждого диагноза на основе
5 консолидации данных из трех источников:
6 - patients.csv - данные о пациентах
7 - medications.xlsx - данные о назначенных лекарствах
8 - drug_prices.json - данные о ценах на лекарства
9
10 ## Установка и запуск
11
12 ### 1. Установка зависимостей
13 ```bash
14
15 • Общая стоимость: 23,847 руб.
16 • Средняя цена: 611 руб.
17 • Пациентов: 31
18
19 [Создание визуализаций]
20
21 Создание визуализаций...
22 ✓ Визуализации сохранены в папке 'results/'
23 - diagnosis_costs_comparison.png
24 - cost_distribution_pie.png
25
26 [Сохранение результатов]
27
28 Сохранение результатов...
29 ✓ Результаты сохранены в папке 'results/'
30 - diagnosis_cost_analysis.xlsx
31 - summary_statistics.json
32
33 =====
34 🎉 Анализ успешно завершен!
```



```
1 # Анализ стоимости лекарств по диагнозам - Вариант 14
2
3 ## Описание
4 Проект для анализа общей стоимости назначенных лекарств для каждого диагноза на основе
5 консолидации данных из трех источников:
6 - patients.csv - данные о пациентах
7 - medications.xlsx - данные о назначенных лекарствах
8 - drug_prices.json - данные о ценах на лекарства
9
10 ## Установка и запуск
11
12 ### 1. Установка зависимостей
13 ```bash
14
15 - cost_distribution_pie.png
16
17 [Сохранение результатов]
18
19 Сохранение результатов...
20 ✓ Результаты сохранены в папке 'results/'
21 - diagnosis_cost_analysis.xlsx
22 - summary_statistics.json
23
24 =====
25 🎉 Анализ успешно завершен!
26
27 📄 ИТОГОВЫЕ РЕЗУЛЬТАТЫ:
28 • Общая стоимость лекарств: 510,148 руб.
29 • Самый затратный диагноз: Гепатит
30 • Самое используемое лекарство: Левотироксин
31 • Все результаты сохранены в папке 'results/'
32
33 ✓ Анализ выполнен успешно!
34 dba@dba-vm: ~/Downloads/var14$
```

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.

