

**МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА  
РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФГБОУ ВО «Кубанский государственный аграрный  
университет имени И.Т.Трубилина»**

**Учетно-финансовый факультет**

**Кафедра экономического анализа**

**ОСНОВЫ АНАЛИЗА ДАННЫХ НА PYTHON**

**Методические указания**

**для контактной и самостоятельной работы  
обучающихся по направлению 38.03.01 Экономика,  
направленность «Бизнес-аналитика»**

**Краснодар  
КубГАУ  
2021**

*Составители: Васильева Н.К., Мороз Н.Ю.*

**Основы анализа данных на Python:** метод. указания для контактной и самостоятельной работы / сост. Васильева Н.К., Мороз Н.Ю.–Краснодар :КубГАУ, 2021. – 25 с.

В методических указаниях для контактной и самостоятельной работы по дисциплине «Основы анализа данных на Python» представлены планы семинарских занятий, темы докладов, вопросы для обсуждения, задачи, тесты, предназначенные для закрепления знаний, умений и навыков.

Предназначены для обучающихся по программе бакалавриата 38.03.01 «Экономика», направленность «Бизнес-аналитика».

Рассмотрены и одобрены методической комиссией учетно-финансового факультета Кубанского государственного аграрного университета, протокол № 9 от 16.06.2021.

Председатель методической комиссии

И.Н. Хромова

© Васильева Н.К.,  
Мороз Н.Ю.,  
составление, 2021  
© ФГБОУ ВО «Кубанский  
государственный аграрный  
университет имени  
И. Т. Трубилина», 2021

## **ВВЕДЕНИЕ**

**Целью** освоения дисциплины «Основы анализа данных на Python» является формирование комплекса знаний о средствах программирования на Python и специализированных библиотек анализа данных, а так же применение полученных знаний для целей бизнес-аналитики.

### **Задачи дисциплины:**

- формирование теоретических знаний и практических навыков по основам анализа данных в Python
- формирование навыков использования Python для работы с табличными данными, импортируемыми или экспортируемыми популярными программами для анализа данных
- приобретение практических навыков использования набора библиотек языка Python для решения задач в сфере бизнес-аналитики

Для освоения представленных материалов рекомендуется ознакомиться с программными вопросами по каждой теме, литературными источниками, нормативами, а также решить практические задачи и оформить результаты каждого этапа работы.

## **Тема 1. Введение в анализ данных на языке Python**

### *План семинара*

1. Python – стандарт для работы с большими данными
2. Прикладные задачи бизнес-аналитиков, для решения которых применяется язык Python

### *Темы докладов:*

1. Преимущества использования Python для анализа данных по сравнению с другими инструментами.
2. Прикладные задачи бизнес-аналитики, для решения которых подходит инструментарий Python.
3. Обзор библиотек и инструментов.
4. Программирование на Python: типы данных, основные структуры, методы и функции, условные операторы, циклы, списковые включения, функция map() и анонимные функции, отладка кода (try/except).
5. Алгоритмы: оптимизация и сложность на примере алгоритмов сортировок.

### *Вопросы для самопроверки*

1. Что такое Python?
2. Каковы преимущества использования Python?
3. Как Python интерпретируется как язык?
4. Что такое PEP 8?
5. Как писать комментарии на Python?
6. Как прокомментировать несколько строк в Python?
7. Что такое строки документации в Python?
8. Обязательны ли отступы в Python?

### *Задачи*

**Задача 1.** Создать следующего вида исходный файл из восьми строк, содержащий в каждой строке значение исходной

системы счисления, результирующей системы счисления и пяти различных целых положительных чисел, корректных для исходной системы счисления:

10 2 121 13 10 2018 5

**Задача 2.** Создать следующего вида исходный файл из десяти строк, содержащий в каждой строке ФИО, дату рождения, дату получения паспорта и баллы ЕГЭ по трём предметам:

Апельсинов А.А. 08.02.2000 17.03.2014 90 100 91

Виноградов В.В. 09.03.1999 15.04.2013 67 99 98

Яблоков Я.Я. 10.04.2000 19.05.2014 94 94 94

Морковкин М.М. 11.05.1999 17.06.2013 91 82 73

**Задача 3.** Не используя готовые сторонние подключаемые функции для факториала, `int()`, `bin()` и т.п., написать программу на языке Python 3.x, которая бы вычисляла среднее значение баллов ЕГЭ, сортировала строки по указанной колонке в обратном порядке (от большего к меньшему) и выводила результат следующего вида (для сортировки по дате рождения):

Яблоков Я.Я. | 10.04.2000 | 19.05.2014 | 94 94 94 -> 94

Апельсинов А.А. | 08.02.2000 | 17.03.2014 | 90 100 91 -> 93,666666

Морковкин М.М. | 11.05.1999 | 17.06.2013 | 91 82 73 -> 82

Виноградов В.В. | 09.03.1999 | 15.04.2013 | 67 99 98 -> 88

### *Тесты:*

1. Какие из приведенных стилей программирования поддерживает язык Python?

а) Процедурный

б) Объектно-ориентированный

- в) Функциональный
- г) Смешанный

2. Каким способом можно объявлять переменные в Python:

- а) `a=5`
- б) `a=int (5)`
- в) `int a=5`

3. Какая функция отвечает за вывод на экран?

- а) `cout<<a`
- б) `out (a)`
- в) `print (a)`

4. Какая функция отвечает за открытие файла?

- а) `file()`
- б) `open()`
- в) `open_file()`

5. В каком из вариантов присутствует ошибка?

- а) `a=5`
- б) `print ('a')`
- в) `while True`
- г) `print(a)`
- д) `a=open("file.txt")`

## Тема 2. Основные операторы и типы данных

### *План семинара*

1. Работа со строками
2. Циклы for
3. Преобразования между типами
4. Работа с последовательностями
5. Работа с текстовыми файлами
6. Работа с модулями

### *Темы докладов:*

1. Библиотека numpy: векторы и массивы, специальные типы данных.
2. Библиотека pandas.
3. Основы работы с датафреймами: загрузка, очистка, фильтрация, группировка и агрегация.
4. Описательные статистики, распределения, разведывательный анализ, работа с пропущенными значениями, постановка гипотезы.
5. Сбор данных из открытых источников: web-scraping, работа с API, парсинг текста и таблиц, сохранение файлов.

### *Вопросы для самопроверки*

1. Что такое функция в Python?
2. Что такое локальные и глобальные переменные в Python?
3. Что такое лямбда-функция?
4. Какие типы данных поддерживаются в Python?
5. Что такое индексы?
6. Что такое отрицательные индексы и почему они используются?
7. Что такое словарь в Python?

## *Задачи*

**Задача 1.** Импортируйте набор данных о пассажирах «Титаника». Проведите разведывательный анализ данных. Поставьте гипотезу. Проведите необходимые преобразования данных.

**Задача 2.** На основании поставленной гипотезы в задании 1 выберите модели для предсказания. Обоснуйте выбор. Постройте выбранные модели, выберите лучшую. Постройте ансамбль моделей. Улучшились ли предсказания? Оформите результаты в виде отчета.

**Задача 3.** Вариант проекта: соберите данные для анализа с сайта Федеральной службы государственной статистики. Проведите предварительную обработку данных. Поставьте задачу классификации. Проведите необходимые преобразования данных. Постройте модель. Оцените, насколько хорошо модель справляется с поставленной задачей. Опишите результат, назовите возможные причины плохой работы модели (в случае, если результаты неудовлетворительные).

## *Тесты*

1. Что делает команда `import`
  - а) импортирует файл модуля
  - б) создает функцию
  - в) удаляет файл
2. Выберите вариант правильного удаления переменной `a`
  - а) `del(a)`
  - б) `delete(a)`
  - в) `delete=a`



3. Какое значение  $1//2$  вернет выражение в среде IDLE?

- а) 0
- б) 0.5
- в) 0.50

4. Как называется встроенный в языке Python тип данных неупорядоченной коллекции из нуля или более пар ключ-значение?

- а) dict
- б) set
- в) list
- г) frozenset

5. Если предположить, что класс Mydict наследует класс dict, то каким класс dict является по отношению к классу Mydict?

- а) дочерним
- б) подклассом
- в) базовым

### **Тема 3. Обмен данными между Python и популярными табличными приложениями**

#### ***План семинара***

1. Формат табличных данных CSV
2. Выгрузка данных для обработки в Python
3. Открытие CSV в Python
4. Формирование файла CSV средствами Python
5. Загрузка сформированных в Python данных в программу
6. Формат табличных данных с фиксированной шириной ячеек (FWF)
7. Прямая интеграция с документами Excel

#### ***Темы докладов:***

1. Виды графиков, их корректное использование и интерпретация.
2. Принципы хорошей визуализации, основные ошибки при визуализации данных.
3. Основы визуализации в matplotlib.
4. Оформление и кастомизация графиков.
5. Визуализации отфильтрованных и сгруппированных данных.

#### ***Вопросы для самопроверки***

1. Что такое лямбда-функция?
2. Какие типы данных поддерживаются в Python?
3. Что такое индексы?
4. Что такое отрицательные индексы и почему они используются?
5. Что такое словарь в Python?
6. Как получить доступ к значениям в словаре?
7. Как получить список всех ключей в словаре?

## Задачи

**Задача 1.** Создайте книгу Excel в Python. Подпишите «шапку» таблицы. Заполните все ячейки одним и тем же числом 123. Добавьте строки с суммой и средним значением.

**Задача 2.** Возьмём три строки:

```
s1 = 'daleonov@specialist.ru'  
s2 = 'not an email'  
s3 = """ sadjflak asdkfjla asklfj daleonov@specialist.ru  
afasdk  
asdflla asf ;asldfka  
fas;dlfka user@gmail.com asklfjalsdf asfk info@mail.ru  
sdfkljasl asldkfjal  
"""
```

Первая, очевидно, почтой не является. Вторая является. Третья - это текст, содержащий адреса электронной почты, но сам по себе он не почта, а текст. Создайте регулярные выражения.

**Задача 3.** Вы хотите узнать на сайте Центробанка РФ курсы доллара с 1 по 31 января 2021 года, на ваш запрос появилась такая ссылка в браузере:  
[http://www.cbr.ru/scripts/XML\\_dynamic.asp?date\\_req1=01/01/2021&date\\_req2=31/01/2021&VAL\\_NM\\_RQ=R01235](http://www.cbr.ru/scripts/XML_dynamic.asp?date_req1=01/01/2021&date_req2=31/01/2021&VAL_NM_RQ=R01235) Это текст, который содержит интересующие вас данные. Какие? Охарактеризуйте их. Как называется данная выгрузка?

## Тесты

1. Что выведет следующий код?

```
d = lambda p: p * 2
```

```
t = lambda p: p * 3
x = 2
x = d(x)
x = t(x)
x = d(x)
print(x)
```

- а) 7
- б) 12
- в) 24
- г) 36
- д) 48

2. Что выведет код `print(type(1 / 2))`, при его исполнении?  
Используется Python 3.x.

- а) class 'int'
- б) class 'number'
- в) class 'float'
- г) class 'double'
- д) class 'tuple'

3. Какие из приведённых утверждений являются ложными?

- а) Программы на Python состоят из модулей.
- б) Переменные в Python автоматически объявляются перед первым использованием.
- в) Python не является интерпретируемым языком программирования.
- г) Переменная - это именованная область жёсткого диска

4. Как называется инструкция, которая определяет данные соответствующие переменной?

- а) Инструкция присваивания
- б) Инструкция цикла

- в) Инструкция определения
- г) Инструкция ветвления

5. Какие значения из указанных могут принимать переменные типа `int`?

- а) 10
- б) 7.8
- в) -10
- г) `'int10'`

## Тема 4. Использование регулярных выражений (regex)

### *План семинара*

1. Описание набора символов
2. Описание количества символов
3. Начало и конец строки
4. Степень детализации регулярных выражений
5. Группы в регулярных выражениях
6. Библиотека для регулярных выражений в Python

### *Темы докладов:*

1. Создание интерактивных визуализаций в Plotly.
2. Нейронные сети. Области применения.
3. Использование существующих решений для собственных задач
4. Американский и европейский стандарт записи дат
5. Формат записи дат ISO

### *Вопросы для самопроверки*

1. Какие типы данных поддерживаются в Python?
2. Что такое индексы?
3. Что такое отрицательные индексы и почему они используются?
4. Что такое словарь в Python?
5. Как получить доступ к значениям в словаре?
6. Как получить список всех ключей в словаре?
7. Что такое pandas в Python?

### *Задачи*

**Задача 1.** Напишите программу Pandas для отображения всех записей файла REGIONS.

**Задача 2.** Напишите программу Pandas для отображения всех идентификаторов местоположений из файла

местоположений.

**Задача 3.** Напишите программу Pandas для извлечения первых 7 записей из файла сотрудников.

**Задача 4.** Напишите программу Pandas, чтобы выбрать отдельный идентификатор отдела из файла сотрудников.

**Задача 5.** Напишите программу Pandas, в которой будут отображаться имя и фамилия, а также номер отдела для всех сотрудников, чья фамилия «Иванов».

### *Тесты*

1. Данные какого типа возвращает функция `input ()`?
  - а) float
  - б) int
  - в) bool
  - г) str
2. Какие из приведённых утверждений правдивы?
  - а) В Python не предусмотрено инструкций для ввода и вывода данных.
  - б) Функция `input ()` возвращает текстовые данные.
  - в) Функция `print ()` не имеет выходных данных.
  - г) Функция `print ()` возвращает числовые данные.
3. Выберите инструкцию, присваивающую переменной `a` значение, введённое пользователем с клавиатуры.
  - а) `print (a)`
  - б) `input (a)`
  - в) `a = input ()`
  - г) `a = print ()`

4. Какие встроенные структуры данных существуют в Python?

- a) frozenset
- б) map
- в) bytes
- г) list
- д) set
- е) array

5. Что будет напечатано при выполнении кода `print(type(1 / 2))`?

- a) type 'int'
- б) type 'number'
- в) type 'double'
- г) type 'tuple'



## Тема 5. Выгрузка данных из онлайн API

### *План семинара*

1. Отправка запросов из Python
2. Формат иерархических данных JSON
3. Формат иерархических данных XML

### *Темы докладов:*

1. Нейронные сети. Области применения.
2. Использование существующих решений для собственных задач
3. Американский и европейский стандарт записи дат
4. Формат записи дат ISO
5. Формат хранения дат Unix Time
6. Регулярные выражения regex
7. Сбор данных с произвольных веб-страниц
8. Автоматизация передачи входных данных через аргументы командной строки
9. Функции построения графиков в pandas
10. Недостатки Python

### *Вопросы для самопроверки*

1. Как получить текущую версию Python?
2. Как можно удалить переменные в Python?
3. Расскажите об интеграции Python с Excel, 1С, SAP
4. Как пользоваться вспомогательными встроенными функциями `zip()` и `enumerate()`?
5. Как происходят преобразования между типами?
6. Расскажите о преобразовании строк в числа и обратно (функции `str`, `int`, `float`)
7. Как получить преобразования одних последовательностей в другие - (функции `list`, `set` и т.п.)?
8. Какие объекты можно обходить в цикле?

## ***Задачи***

**Задача 1.** Напишите программу Pandas для отображения имени, фамилии, оклада и идентификатора менеджера, где идентификаторы менеджера равны нулю.

**Задача 2.** Напишите программу Pandas для отображения имени, фамилии, оклада и идентификатора менеджера, где идентификаторы менеджера не равны нулю.

**Задача 3.** Напишите программу Pandas для создания и отображения логического ряда, где True для not null и False для null-значений или пропущенных значений в столбце state\_province файла местоположений.

## ***Тесты***

1. Какие из приведённых утверждений являются ложными?

- а) Программы на Python состоят из модулей.
- б) Переменные в Python автоматически объявляются перед первым использованием.
- в) Python не является интерпретируемым языком программирования.
- г) Переменная - это именованная область жёсткого диска

2. Как называется инструкция, которая определяет данные соответствующие переменной?

- а) Инструкция присваивания
- б) Инструкция цикла
- в) Инструкция определения
- г) Инструкция ветвления

3. Какие значения из указанных могут принимать

переменные типа int?

- а) 10
- б) 7.8
- в) -10
- г) 'int10'

4. Данные какого типа возвращает функция input ()?

- а) float
- б) int
- в) bool
- г) str

5. Какие из приведённых утверждений правдивы?

а) В Python не предусмотрено инструкций для ввода и вывода данных.

- б) Функция input () возвращает текстовые данные.
- в) Функция print () не имеет выходных данных.
- г) Функция print () возвращает числовые данные.

## Тема 6. Обзор модулей Python для анализа данных

### *План семинара*

1. Модуль для работы с числовыми данными NumPy
2. Модуль для работы с табличным представлением данных  
Pandas
3. Обзор библиотеки Matplotlib для визуализации данных  
в Python

### *Темы докладов:*

1. Модуль для работы с числовыми данными NumPy
2. Особенные типы данных в NumPy
3. Работа с векторами и матрицами
4. Вычисление главных статистических метрик с помощью  
NumPy
5. Модуль для работы с табличным представлением данных  
Pandas
6. Преобразование словарей в табличный формат Pandas
7. Загрузка данных из внешних источников
8. Инструменты визуализации данных для Python
9. Обзор библиотеки Matplotlib

### *Вопросы для самопроверки*

1. Что такое текстовые и бинарные файлы?
2. Что такое кодировки?
3. Как сделать перенос строк?
4. Как открывать нетекстовые файлы?
5. Как производить хранение дат в программах?
6. Назовите основные модули для работы с датами в Python
7. Каков формат табличных данных CSV?
8. Как выгрузить данные для обработки в Python?

### *Задачи*

**Задача 1.** Напишите программу Pandas для отображения имени, фамилии, оклада и номера отдела для тех сотрудников,

менеджеры которых имеют ID 120, 103 или 145.

**Задача 2.** Напишите программу Pandas, чтобы отобразить имя, фамилию, оклад и номер отдела для тех сотрудников, для которых буква п является третьим символом в их имени.

**Задача 3.** Напишите программу Pandas для отображения имени, идентификатора задания, зарплаты и отдела для тех сотрудников, которые не работают в отделах 50, 30 и 80.

### *Тесты*

1. Какие встроенные структуры данных существуют в python?
  - а) frozenset
  - б) map
  - в) bytes
  - г) list
  - д) set
  - е) array
2. Что будет напечатано при выполнении кода `print(type(1 / 2))`?
  - а) type 'int'
  - б) type 'number'
  - в) type 'double'
  - г) type 'tuple'
3. Какое значение `1//2` вернет выражение в среде IDLE?
  - а) 0
  - б) 0.5
  - в) 0.50

4. Выберите инструкцию, присваивающую переменной a значение, введённое пользователем с клавиатуры.

- а) `print (a)`
- б) `input (a)`
- в) `a = input ()`
- г) `a = print ()`

5. Какая функция отвечает за вывод на экран?

- а) `cout<<a`
- б) `out (a)`
- в) `print (a)`

## Список рекомендованной литературы

1. Блэнд, Д. Тестирование бизнес-идей / Дэвид Блэнд, Алекс Остервальдер ; пер. с англ. - Москва : Альпина Паблишер, 2020. - 354 с. - Режим доступа: <https://znanium.com/catalog/product/1222476>

2. Глебов, В. И. Практикум по математической статистике. Проверка гипотез с использованием Excel, MatCalc, R и Python : учебное пособие / В. И. Глебов, С. Я. Криволапов. — Москва : Прометей, 2019. — 86 с. - Режим доступа: <https://www.iprbookshop.ru/94504.html>

3. Гоулман, Д. Управление бизнесом / Гоулман Д., Друкер П., Дэвенпорт Т. - Москва : Альпина Пабли., 2016. - 290 с. (Harvard Business Review 10 лучших статей) ISBN 978-5-9614-5860-2. - Текст : электронный. - URL:

<https://znanium.com/catalog/product/701800> 1. Маккинли, Уэс Python и анализ данных / Уэс Маккинли; перевод А. Слинкина. — 2-е изд. — Саратов : Профобразование, 2019. — 482 с. - Режим доступа: <https://www.iprbookshop.ru/88752.html>

4. Гуриков, С. Р. Основы алгоритмизации и программирования на Python: учебное пособие / С.Р. Гуриков. — Москва: ФОРУМ: ИНФРА-М, 2020. — 343 с. — (Высшее образование: Бакалавриат). - Режим доступа: <https://znanium.com/catalog/product/1206074>

5. Кон, М. Agile. Оценка и планирование проектов: Практическое руководство / Кон М. - М.: Альпина Паблишер, 2018. - 418 с. - Режим доступа: <https://znanium.com/catalog/product/1003486>

6. Коттер, Д. Управление изменениями: Практическое руководство - М.: Альпина Паблишер, 2016. - 226 с.: Режим доступа: <https://znanium.com/catalog/product/1003136>

7. Программные системы статистического анализа. Обнаружение закономерностей в данных с использованием системы R и языка Python : учебное пособие / В. М. Волкова, М. А. Семёнова, Е. С. Четвертакова, С. С. Вожов. — Новосибирск:

Новосибирский государственный технический университет, 2017. — 74 с. - Режим доступа: <https://www.iprbookshop.ru/91682.html>

8. Свод знаний по управлению бизнес-процессами. BPM СВОК 3.0: Учебное пособие / Под ред. Белайчук А.А. - Москва :Альпина Пабли., 2016. - 480 с. Режим доступа: <https://znanium.com/catalog/product/558829>

9. Сузи, Р. А. Язык программирования Python : учебное пособие / Р. А. Сузи. — 3-е изд. — Москва : Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2020. — 350 с. - Режим доступа: <https://www.iprbookshop.ru/97589.html>

10. Чиркова, Е. В. Как оценить бизнес по аналогии: Пособие по использованию сравнительных рыночных коэффициентов Учебное пособие / Чиркова Е.В., - 4-е изд., исправ. и доп. - М.:Альпина Паблишер, 2017. - 284 с. Режим доступа: <https://znanium.com/catalog/product/1002519>.

11. Шелудько, В. М. Язык программирования высокого уровня Python. Функции, структуры данных, дополнительные модули : учебное пособие / В. М. Шелудько. — Ростов-на-Дону, Таганрог : Издательство Южного федерального университета, 2017. — 107 с. - Режим доступа: <https://www.iprbookshop.ru/87530.html>



# **ОСНОВЫ АНАЛИЗА ДАННЫХ НА PYTHON**

*Методические указания*

*Составители:*

**Васильева Надежда Константиновна**  
**Мороз Наталья Юрьевна**

В авторской редакции

Подписано в печать \_\_\_\_\_ 2021. Формат бумаги 60 × 84<sup>1</sup>/<sub>8</sub>  
Усл. печ. л. – 2,0. Уч.-изд. л. – 1,6.

Кубанский государственный аграрный университет.  
350044, г. Краснодар, ул. Калинина, 13