
**ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ГОРОДА МОСКВЫ ДОПОЛНИТЕЛЬНОГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
ЦЕНТР ПРОФЕССИОНАЛЬНЫХ КВАЛИФИКАЦИЯ И СОДЕЙСТВИЯ ТРУДОУСТРОЙСТВУ
«ПРОФЕССИОНАЛ»**

УТВЕРЖДАЮ

Первый заместитель директора
ГБОУ ДПО Центр «Профессионал»

И. В. Курнева
«14» февраля 2022 г.



**КОНТРОЛЬНЫЕ ВОПРОСЫ
для подготовки к итоговой аттестации**

«Аналитик данных»

профессиональная переподготовка (дополнительное профессиональное образование)
(наименование программы)

Теоретические вопросы:

1. Типы данных в Python.
2. Предобработка данных.
3. Шкалы измерений.
4. Списки в Python. Работа со списками.
5. Кортежи.
6. Аномалии.
7. Вложенные списки.
8. Словари в Python.
9. Функции. Возвращение значений из функций.
10. Классификация и кластеризация – суть и назначение.
11. Функции как объект. Лямбда-функции.
12. Рекурсия.
13. Работа с пропусками.
14. Категоризация данных.
15. Методы в Python.
16. Шкалы измерений.
17. Условный оператор.
18. Строки в Python.
19. Множества в Python.

20. Чем отличаются ошибки первого и второго рода
21. Модули и пакеты.
22. Генераторы.
23. Корреляция и регрессионный анализ.
24. Генеральная совокупность и выборка.
25. Постановка задачи регрессионного анализа.
26. Шкалы измерений.
27. Модуль random.
28. Отношения между таблицами.
29. Агрегирование данных.
30. Визуализация данных.
31. Библиотека Numpy в анализе данных.
32. Вычислительные возможности Numpy.
33. Массивы в Numpy.
34. Библиотека Pandas.
35. Объект Pandas. Ser
36. Объект Pandas. DataFrame.

Практические задания:

1. Повторите 1000 раз следующий эксперимент: сгенерируйте две матрицы размера 10 на 10 из стандартного нормального распределения, перемножьте их (как матрицы) и найдите максимальный элемент. Какое среднее значение по экспериментам у максимальных элементов?
2. Сгенерируйте 1000 чисел из распределения Пуассона с параметром $\lambda=5$. Сформируйте массив, в котором в i -м элементе будет записано количество сгенерированных чисел, равных i .
3. Реализуйте функцию, которая принимает на вход numpy-массив целых чисел a , и генерирует массив, в котором число i встречается $a[i]$ раз.
4. Напишите функцию $scale(X)$, которая принимает на вход матрицу и масштабирует каждый ее столбец (вычитает среднее и делит на стандартное отклонение). Убедитесь, что в функции не будет происходить деления на ноль. Протестируйте на каких-нибудь данных.¶
5. Сгенерируйте матрицу с элементами из нормального распределения $N\{10, 1\}$. Найдите ее:
 - определитель
 - след
 - наименьший и наибольший элементы
 - спектральную норму
 - норму Фробениуса
 - собственные числа
 - обратную матрицу

6. Напишите функцию `remove_duplicates(a)`, которая принимает на вход список и возвращает его же, но без дубликатов и в отсортированном виде. Например, для списка `['ночь', 'улица', 'фонарь', 'аптека', 'аптека', 'улица', 'фонарь']` результат должен быть `['аптека', 'ночь', 'улица', 'фонарь',]`.
7. Напишите функцию `sort_tokens(s)`, которая принимает на вход строку, разделяет ее на токены по запятым, сортирует токены с помощью функции `sorted()` и снова соединяет их в одну строку через запятые. Например, для строки `"ночь,фонарь,улица"` функция должна выдать `"ночь,улица,фонарь"`. Придумайте несколько входных строк и протестируйте функцию на них.¶
8. Необходимо создать матрицу 5 x 5 с 1,2,3,4 под диагональю.
9. Создайте массив 3 x 3 x 3 со случайными значениями.
10. Посчитайте размерность данных для датафрейма

```
df = pd.DataFrame({'float': [1.0],  
                  'int': [1],  
                  'datetime': [pd.Timestamp('20180310')],  
                  'string': ['foo']})  
df.dtypes
```
11. Выведите тип данных каждой переменной для датафрейма

```
df = pd.DataFrame({'float': [1.0],  
                  'int': [1],  
                  'datetime': [pd.Timestamp('20180310')],  
                  'string': ['foo']})  
df.dtypes
```
12. Необходимо создать вектор со значениями от 10 до 49.
13. Импортируйте библиотеку NumPy под именем np.
14. Создайте объект Series состоящий из: [1, 1, 2, 3, 5, 8].
15. Создайте DataFrame

int_col	text_col	float_col
0	1 alpha	0.00
1	2 beta	0.25
2	3 gamma	0.50
3	4 delta	0.75
4	5 epsilon	1.00
16. Вычислите массив `a_centered`, отняв от значений массива `a` средние значения соответствующих признаков, содержащиеся в массиве `mean_a`. Вычисление должно производиться в одно действие. Получившийся массив должен иметь размер 5x2.
17. Найдите скалярное произведение столбцов массива `a_centered`. В результате должна получиться величина `a_centered_sp`. Затем поделите `a_centered_sp` на N-1, где N - число наблюдений.
18. Создайте массив Numpy под названием `a` размером 5x2, то есть состоящий из 5 строк и 2 столбцов. Первый столбец должен содержать числа 1, 2, 3, 3, 1, а второй - числа 6, 8, 11, 10, 7. Будем считать, что каждый столбец - это признак, а строка - наблюдение. Затем найдите среднее значение по каждому признаку, используя метод `mean` массива Numpy. Результат запишите в массив `mean_a`, в нем должно быть 2 элемента.