Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
Санкт-Петербургский горный университет

Кафедра информатики и компьютерных технологий

ИНФОРМАТИКА. СТАТИСТИЧЕСКАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Методические указания по выполнению лабораторных работ для студентов бакалавриата направления 21.03.01

САНКТ-ПЕТЕРБУРГ
2021

УДК 519.221 (073)

ИНФОРМАТИКА. СТАТИСТИЧЕСКАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ: Методические указания по выполнению лабораторных работ / Санкт-Петербургский горный университет. Сост.: *С.Б. Крыльцов, М.А. Коробицына.* СПб, 2021. 33 с.

В методических указаниях представлены лабораторные работы, которые позволяют выработать навыки обработки массивов экспериментальных данных с применением различных программных средств – табличного процессора Excel, языка программирования Python без дополнительных библиотек, а также модулем pandas, предназначенным для статистической обработки данных в Python. Каждая лабораторная работа сопровождается графическими материалами и листингами, содержащими исходный код на языке Python, наглядно демонстрируя реализацию подходов к обработке данных.

Научный редактор доц. *А.Б. Маховиков*

Рецензент канд. техн. наук *К.В. Столяров* (Telum Inc.)

ВВЕДЕНИЕ

Подготовка будущих инженеров в техническом вузе неразрывно связана с выполнением студентами различных практических работ и прохождением производственной практики, в ходе которых обучающиеся не только получают соответствующие направлению подготовки навыки, но и накапливают значительный объем материала, полученного в ходе экспериментов. Как правило, экспериментальные данные представлены в виде набора дискретных значений, который в дальнейшем необходимо обрабатывать в специализированном программном обеспечении.

В настоящее время существует широкий выбор программного обеспечения для статистической обработки данных. Несмотря на то, что такое программное обеспечение предоставляет множество функций, которые в один-два клика позволяют вычислить автокорреляцию сигнала, его спектр, получить распределение случайной величины, визуализировать ряд данных и пр., понимание того, что происходит внутри такого программного обеспечения, является одним из ключевых навыков в статистике.

В данных методических указаниях приведены лабораторные работы, цель которых – выполнить статистическую обработку данных в табличном процессоре Excel, а затем реализовать такую же обработку средствами языка Python без дополнительных библиотек и сравнить эту реализацию с выполнением статистической обработки данных с помощью специализированного модуля pandas для Python.

ТРЕБУЕМОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

Для выполнения лабораторных работ из данных методических указаний минимально необходимое программное обеспечение включает в себя: интерпретатор языка Python версии 3.6 и выше, интерактивную вычислительную среду на основе веб-интерфейса Jupyter Notebook для Python3, программную библиотеку для обработки и анализа данных pandas для Python3, Microsoft Excel версии 2007 и выше. Для упрощения процесса установки необходимых библиотек для Python целесообразно установить Anaconda – свободно распространяемый набор программного обеспечения для статистической обработки данных, машинного обучения, формирования аналитики с использованием средств языков программирования Python и R. Для установки Anaconda под операционной системой (ОС) Windows необходимо загрузить и выполнить инсталлятор с официального сайта https://www.anaconda.com/ выбрав версию релиза для индивидуального использования – «Individual Edition» – под 32- либо 64-разрядную архитектуру, в зависимости от архитектуры ОС.

При установке Anaconda рекомендуется использовать параметры, представленные на рисунке 1, во избежание конфликтов с другими версиями интерпретатора Python, установленными в системе.

В методических указаниях используется следующие версии программного обеспечения:

1. Anaconda 3 (версия 2020.02).
2. Excel 2019 (версия16.0.6742.2048).

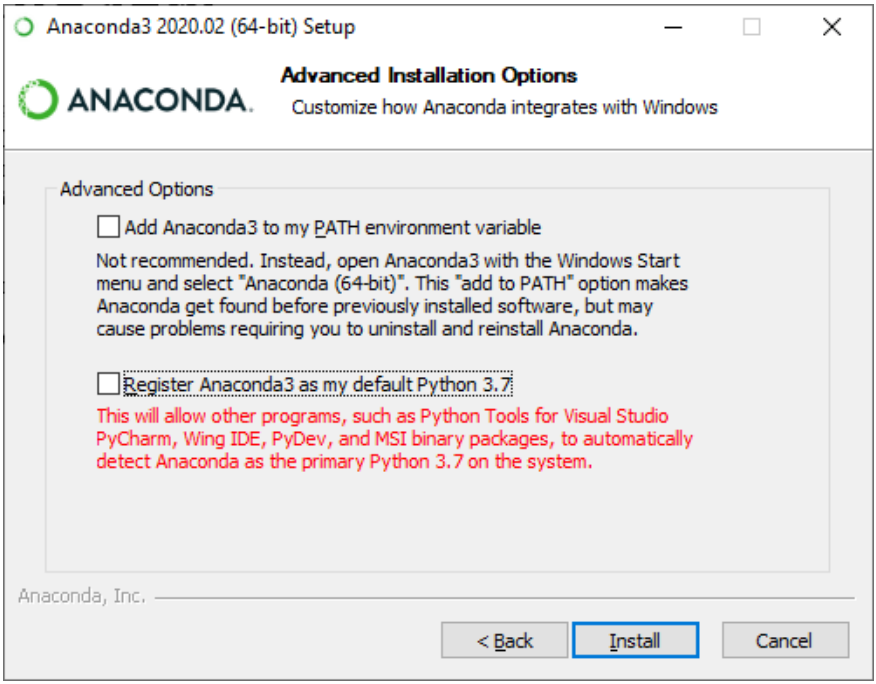


Рис. 1 – Рекомендуемые параметры для установки Anaconda 3.

ЛАБОРАТОРНАЯ РАБОТА 1. ПОДГОТОВКА И ИМПОРТИРОВАНИЕ ДАННЫХ

ЗАДАНИЕ

Импортировать и вывести содержимое набора данных «Proved Oil Reserves: 2010» в Excel, Python, Pandas.

ОБЩИЕ СВЕДЕНИЯ

Для обработки данных необходимо сначала эти данные импортировать. Одним из наиболее распространённых форматов данных для переноса между устройствами является формат CSV (Comma-separated values) – формат, в котором табличные данные пред-

ставлены построчно с разделением значений, соответствующих колонкам, через запятую.

Данные, сохранённые в CSV, обычно представлены текстовыми файлами с расширением .csv, которые можно открыть в любом текстовом редакторе. Принципы структуризации данных в .csv файлах следующие:

1. В первой строке через запятую записываются названия столбцов в кавычках.
2. Во второй и далее строках записываются значения, соответствующие каждому столбцу, также через запятую.
3. В конце строк не ставятся запятые.
4. Если значение отсутствует – оно не заполняется, но всё равно отделяется запятой. Названия столбцов можно также пропускать.

Рассмотрим работу с CSV данными на примере набора данных «Proved Oil Reserves: 2010», свободно распространяемого на сайте Open Energy Information, доступным для загрузки по ссылке:

https://openei.org/datasets/dataset/proved-oil-reserves-2010

Набор представляет собой оценку подтверждённых государственных запасов сырой нефти на начало 2010-го года. Данные сохранены в CSV-файле «provedoilreserves.csv». Содержимое файла представлено двумя столбцами: страной и количеством сырой нефти в баррелях. Пример строки из CSV-файла:

"Saudi Arabia","264,600,000,000"

РЕШЕНИЕ В EXCEL

Для начала следует подготовить данные для загрузки. Открыв CSV-файл в любом текстовом редакторе, можно отметить две особенности: в файле нет названий столбцов, а численное значение представлено строкой с отделением групп чисел запятой.

Для упрощения дальнейшей работы следует добавить названия столбцов, добавив в начало файла строку:

”Страна”, “Запасы сырой нефти”

В Excel значения CSV можно добавить, воспользовавшись меню «Data» – «From Text/CSV» (рисунок 2).

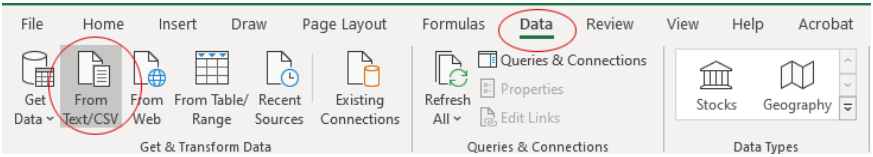


Рис. 2 – импортирование CSV файла в Excel.

В раскрывающемся списке File Origin следует выбрать кодировку нелатинских символов, если они содержатся в файле .csv. Для файлов, содержащих кириллицу, чаще всего кодировка будет 65001: Unicode либо 1251: Cyrillic.

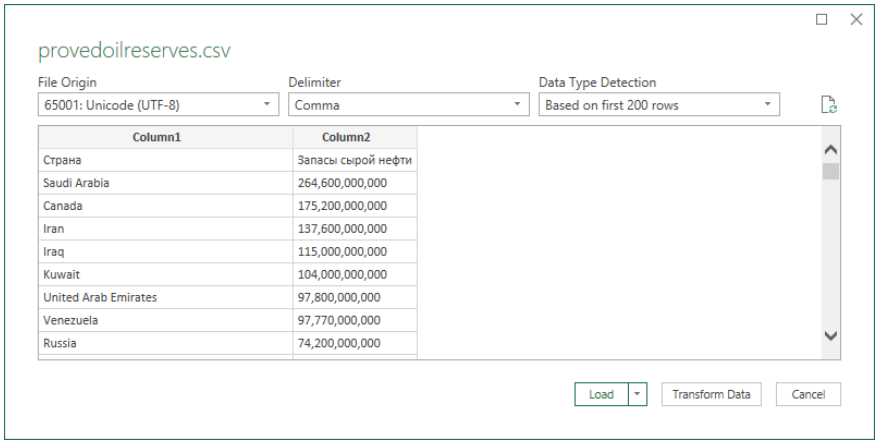


Рис. 3 – Выбор параметров импорта CSV файла в Excel.

В файлах .csv (несмотря на название) в качестве разделителя могут использоваться другие символы кроме запятой. В таком случае в раскрывающемся списке «Delimiter» следует выбрать символ разделителя.

Функция «Data type detection» отвечает за тип, который будет присвоен ячейкам в Excel, содержащим данные после импорта их из файла .csv. Если .csv файл содержит несколько таблиц данных, то целесообразно выбирать пункт «Based on entire dataset», в противном случае – «Based on first 200 rows». При больших наборах данных опция «Based on entire dataset» может существенно увеличить время их импорта на лист Excel.

Из рисунка 1 видно, что Excel обработал первую строку как данные, что обусловлено использованием строковых значений при записи запасов сырой нефти вместо числовых. Чтобы принудительно использовать первую строку в качестве заголовков столбцов необходимо нажать на кнопку «Transform Data», затем «Use First Row as Headers», как показано на рисунке 4.

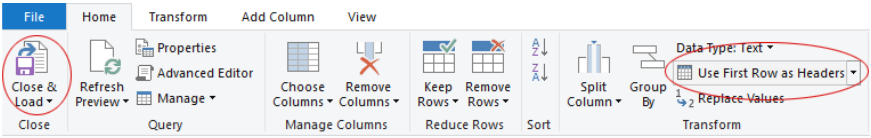


Рис. 4 – Использование первой строки в качестве наименований столбцов.

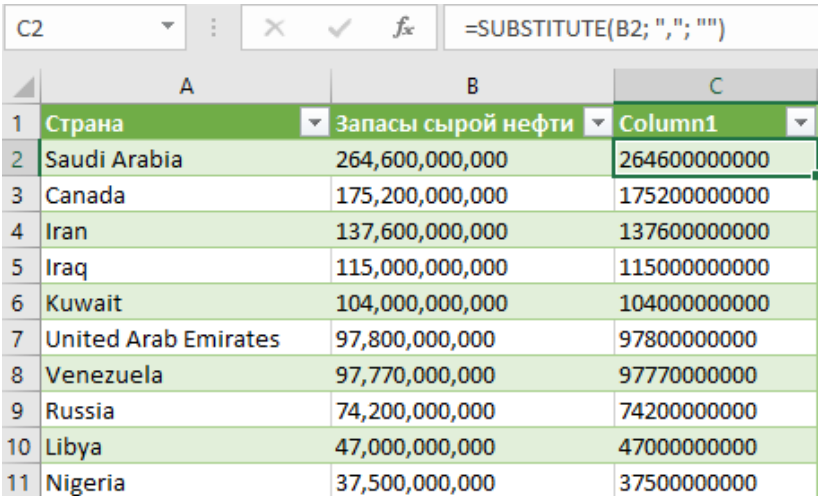
После нажатия на кнопку «Close and Load» импортированные в Excel данные должны появиться на листе Excel как показано на рисунке 5.

	A	B
1	Страна	Запасы сырой нефти
2	Saudi Arabia	264,600,000,000
3	Canada	175,200,000,000
4	Iran	137,600,000,000
5	Iraq	115,000,000,000
6	Kuwait	104,000,000,000
7	United Arab Emirates	97,800,000,000
8	Venezuela	97,770,000,000
9	Russia	74,200,000,000
10	Libya	47,000,000,000
11	Nigeria	37,500,000,000

Рис. 5 – Набор данных после импорта файла .csv в Excel.

Наличие заголовков таблицы данных в файле .csv позволяет Excel автоматически сформировать фильтры для данных, доступные при нажатии на кнопку  заголовках столбцов. Далее необходимо убрать запятые из столбца «Запасы сырой нефти» с помощью функции *SUBSTITUTE*, которая принимает 3 аргумента: целевую ячейку; строку, которую необходимо заменить; строку, которую необходимо подставить. Для этого в ячейку C2 введём строку «`=SUBSTITUTE(B2; ","; "")`». После протягивания ячейки вниз, мы получим столбец с числовыми данными (рисунок 6), по

которым можно вести дальнейшие расчёты, либо заменить ими столбец «Запасы сырой нефти».



	A	B	C
1	Страна	Запасы сырой нефти	Column1
2	Saudi Arabia	264,600,000,000	264600000000
3	Canada	175,200,000,000	175200000000
4	Iran	137,600,000,000	137600000000
5	Iraq	115,000,000,000	115000000000
6	Kuwait	104,000,000,000	104000000000
7	United Arab Emirates	97,800,000,000	97800000000
8	Venezuela	97,770,000,000	97770000000
9	Russia	74,200,000,000	74200000000
10	Libya	47,000,000,000	47000000000
11	Nigeria	37,500,000,000	37500000000

Рис. 6 – Новый столбец с числовыми данными, полученными заменой запятых в столбце «Запасы сырой нефти».

Однако, значения в новом столбце всё ещё обрабатываются Excel как текст. Для приведения их к числовому формату необходимо для данного столбца применить операцию «Text to Columns» из вкладки «Data», как это показано на рисунке 7. Итоговая таблица должна принять вид, показанный на рисунке 8.

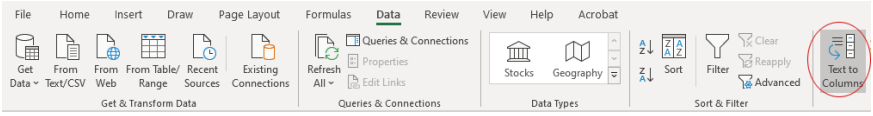


Рис. 7 – Новый столбец с числовыми данными, полученными заменой запятых в столбце «Запасы сырой нефти».

	A	B
1	Страна	Запасы сырой нефти
2	Saudi Arabia	2,646E+11
3	Canada	1,752E+11
4	Iran	1,376E+11
5	Iraq	1,15E+11
6	Kuwait	1,04E+11
7	United Arab Emirates	97800000000
8	Venezuela	97770000000
9	Russia	74200000000
10	Libya	47000000000
11	Nigeria	37500000000

Рис. 8 – Итоговый вид таблицы набора данных.

РЕШЕНИЕ В PYTHON

Для того, чтобы работать с CSV данными в Python, необходимо импортировать модуль *csv*. Код для открытия файла .csv приведён в листинге 1:

```
1 import csv
2 values = []
3 with open('provedoilreserves.csv', encoding='utf-8') as f:
4     csv_reader = csv.DictReader(f, delimiter=',')
5     for row in csv_reader:
6         values.append(row)
7 print(values)
```

Листинг 1 – Код для вывода содержимого .csv файла в консоль Python

Для открытия файла используется стандартная функция Python open, для которой в качестве входных параметров необходимо указать путь к файлу и кодировку в случае необходимости. Для чтения содержимого CSV-файла используется объект *csv.DictReader*,

который в качестве обязательного параметра инициализации принимает переменную, ссылающуюся на открытый файл. Опционально можно выбрать разделитель, передав параметр *delimiter*.

Данный код целесообразно обернуть в блок *with*, что приведёт к автоматическому закрытию файла после выполнения всего кода блока. Объект *csv.DictReader* представляет собой итератор по содержимому файла, возвращающий каждую строку в виде сортируемого словаря с ключами, извлечёнными из первой строки файла, содержащей названия столбцов. Передав все строки в список, можно вывести его на экран командой *print()*:

```
[OrderedDict([('Страна', 'Saudi Arabia'), ('Запасы сырой
нефти', '264,600,000,000')]), OrderedDict([('Страна',
'Canada'), ('Запасы сырой нефти', '175,200,000,000')]), Or-
deredDict([('Страна', 'Iran'), ...
```

Из вывода команды *print()* видно, что значения столбца «Запасы сырой нефти» также представляют из себя строки. Заменить их можно с помощью метода строки *str.replace()* и преобразования полученного после замены запятых значений к целочисленному типу следующим, как показано в листинге 2.

```
1 for value in values:
2     value['Запасы сырой нефти'] = int(value['Запасы сырой
  нефти'].replace(',', ''))
3 print(values)
```

Листинг 2 – Код для преобразования строчных значений столбца «Запасы сырой нефти» к целочисленным значениям

После выполнения кода из листинга 2 вывод должен быть следующим:

```
[OrderedDict([('Страна', 'Saudi Arabia'), ('Запасы сырой
нефти', 264600000000)]), OrderedDict([('Страна', 'Canada'),
```

```
('Запасы сырой нефти', 175200000000)]),
OrderedDict([('Страна', 'Iran'),...
```

РЕШЕНИЕ В PANDAS

В pandas табличные данные обрабатываются в объектах класса DataFrame. Объект DataFrame можно создать на основе встроенной в Pandas функции чтения CSV-файлов *pandas.readcsv()*, которая на вход принимает путь к CSV-файлу. Разделитель данных в CSV-файле настраивается опциональным параметром *sep*, а кодировка файла параметром *encoding*.

```
1 import pandas as pd
2 df = pd.read_csv('provedoilreserves.csv', sep=',',
  encoding='utf-8')
3 df.head()
```

Листинг 3 – Код для считывания файла «provedoilreserves.csv» в DataFrame pandas

Вывод первых строк считанного набора данных можно осуществить вызовом метода *DataFrame.head()*. При выполнении данного кода в Jupyter Notebook вывод должен быть как на рисунке 9.

	Страна	Запасы сырой нефти
0	Saudi Arabia	264,600,000,000
1	Canada	175,200,000,000
2	Iran	137,600,000,000
3	Iraq	115,000,000,000
4	Kuwait	104,000,000,000

Рис. 9 – Вывод первых пяти столбцов DataFrame в Jupyter Notebook.

Значения столбца «Запасы сырой нефти» в DataFrame также необходимо преобразовать к числовым. Для этого применим ано-

нимную функцию к каждой строке DataFrame с помощью метода *DataFrame.apply()* с параметром *axis=1*, что обозначает индексацию по строкам. Значение *axis=0*, которое параметр имеет по умолчанию, приводит к применению функции с индексацией элементов по столбцам. Из каждой строки x, переданной анонимной функции, извлечём строчное значение, соответствующее столбцу «Запасы сырой нефти», в котором удалим все запятые с помощью оператора *replace()*, как это было сделано в решении для Python.

```
df['Запасы сырой нефти'] = df.apply(lambda x:
int(x['Запасы сырой нефти'].replace(',', '')),
axis=1)
```

Листинг 4 – Код для преобразования строчных значений столбца «Запасы сырой нефти» к целочисленным значениям

Результаты вывода метода *DataFrame.head()* после выполнения вышеприведённого кода представлены на рисунке 10.

	Страна	Запасы сырой нефти
0	Saudi Arabia	264600000000
1	Canada	175200000000
2	Iran	137600000000
3	Iraq	115000000000
4	Kuwait	104000000000

Рис. 10 – Вывод первых пяти столбцов DataFrame после преобразования строчных значений столбца «Запасы сырой нефти» к числовым.

ЛАБОРАТОРНАЯ РАБОТА 2. ОСНОВЫ РАБОТЫ С ТАБЛИЧНЫМИ ДАННЫМИ

ЗАДАНИЕ

1. Вычислить количество стран в наборе данных.
2. Вычислить среднее значение подтверждённых запасов сырой нефти, приходящееся на одну страну.
3. Вычислить количество стран с нулевыми подтверждёнными запасами сырой нефти.
4. Отсортировать набор данных по алфавиту.

РЕШЕНИЕ В EXCEL

Для вычисления количества стран следует воспользоваться функцией *COUNT()*, которая считает заполненные ячейки в заданном диапазоне:

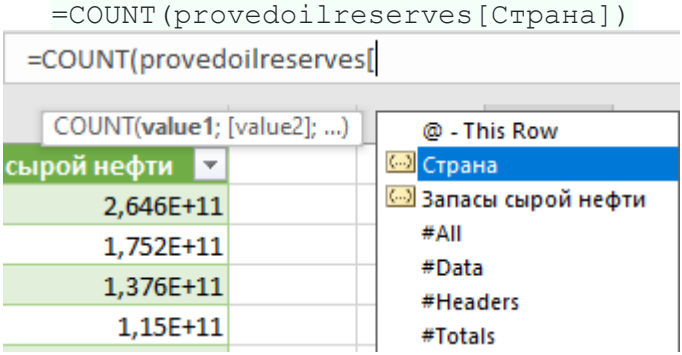


Рис. 11 – Вывод первых пяти столбцов DataFrame в Jupyter Notebook.

Для быстрого доступа к данным столбцов удобно использовать всплывающие подсказки Excel (рисунок 11). Вместо диапазона ячеек либо создания именных диапазонов можно воспользоваться именованным набором данных, который был автоматически создан Excel при импорте CSV-файла. Название набора данных по умолча-

нию присваивается в соответствии с именем CSV-файла. Ссылка на конкретный столбец доступна при заключении его названия в квадратные скобочки.

По умолчанию функция COUNT() считает только ячейки с числовыми значениями, поэтому можно либо применить её к столбцу «Запасы сырой нефти» либо воспользоваться функцией COUNTIF(), которая кроме диапазона данных в качестве второго параметра принимает условие, которому должны соответствовать ячейки для подсчёта. Если передать в качестве условия оператор «*», функция COUNTIF() считает ячейки, заполненные любыми данными:

```
=COUNTIF(provedoilreserves[Страна]; "*")
```

Количество стран с нулевыми подтверждёнными запасами сырой нефти можно вычислить также с помощью функции COUNTIF(), передав в качестве первого параметра столбец «Запасы сырой нефти» и в качестве второго аргумента условие «=0».

Среднее значение подтверждённых запасов сырой нефти, приходящееся на одну страну можно вычислить с помощью функции AVERAGE(), передав ей в качестве параметра столбец «Запасы сырой нефти»:

```
=AVERAGE(provedoilreserves[Запасы сырой нефти])
```

Для сортировки данных в Excel целесообразно воспользоваться функцией сортировки, встроенной в фильтры данных. Для этого необходимо нажать на кнопку  в наименовании столбца «Страна», после чего выбрать пункт «Sort A to Z», как это показано на рисунке 12.

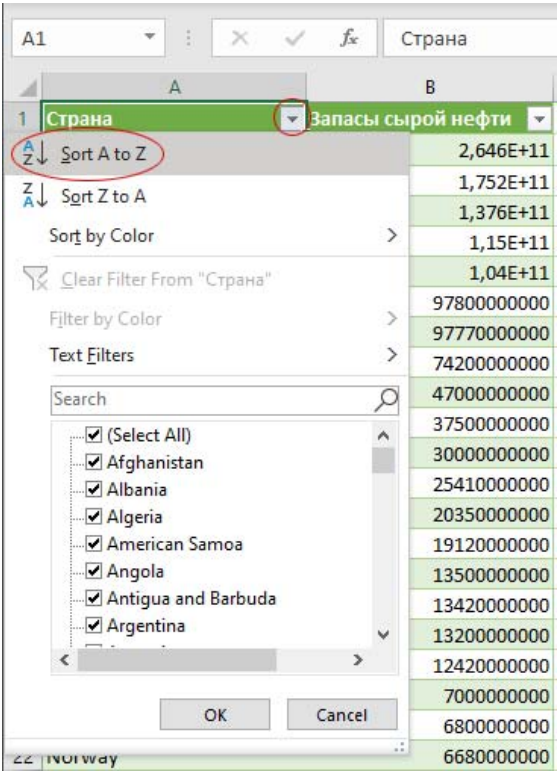


Рис. 12 – Сортировка строк таблицы по столбцу «Страна» по алфавиту.

После применения функций COUNTIF() и AVERAGE() к ячейкам E2, F2 и G2 соответственно, полученные значения должны принять вид, представленный на рисунке 13.

РЕШЕНИЕ НА PYTHON

Количество стран в наборе данных можно получить, посчитав длину списка *values* с помощью функции *len()*. Код для вычисления количества стран представлен в листинге 5.

E	F	G
Количество стран	Количество стран с нулевыми запасами сырой нефти	Средний объём запасов сырой нефти на страну
204	106	6825789461

Рис. 13 – Вычисленное количество стран в наборе данных, количество стран, обладающих нулевым подтверждённым запасов сырой нефти и средний объём запасов сырой нефти в пересчёте на одну страну

```
1 print(len(values))
```

Листинг 5 – Код для преобразования строчных значений столбца «Запасы сырой нефти» к целочисленным значениям

Вывод результатов исполнения кода:

```
204
```

Количество стран в наборе данных, в которых подтверждённый объём запасов сырой нефти равен нулю, можно вычислить с помощью генератора списка (list comprehension), который формирует список с использованием цикла for in и условного оператора if, обёрнутых в квадратные скобки. Код для генерации списка стран с нулевыми подтверждёнными запасами сырой нефти и вывода его длины представлен в листинге 6.

```
1 print(len([row for row in values if row['Запасы сырой
  нефти'] == 0]))
```

Листинг 6 – Код для генерации списка стран с нулевыми подтверждёнными запасами сырой нефти и вывода его длины

Вывод результатов исполнения кода:

```
106
```

Для сортировки стран по алфавиту следует использовать функцию *sort()*, которой в качестве первого параметра необходимо передать список словарей со строками таблицы набора данных, а в качестве второго параметра список ключей, по которым будем производиться сортировка. Для получения списка ключей из значений столбца «Страна» возможно использовать анонимную функцию. Код приведён в листинге 7.

```
1 print(sorted(values, key=lambda k: k['Страна']))
```

Листинг 7 – Код для генерации списка стран с нулевыми подтверждёнными запасами сырой нефти и вывода его длины

Вывод результатов исполнения кода:

```
[OrderedDict([('Страна', 'Afghanistan'), ('Запасы сырой нефти', 0)]), ...
```

Среднее значение подтверждённых запасов сырой нефти можно получить, вычислив сумму всех значений столбца «Запасы сырой нефти» и поделив его на количество строк в таблице. Для этого также целесообразно воспользоваться генератором списка с извлечением значения столбца «Запасы сырой нефти», как это показано в листинге 8.

```
1 rows = [row['Запасы сырой нефти'] for row in values]
2 print(sum(rows)/len(rows))
```

Листинг 8 – Код для генерации списка стран с нулевыми подтверждёнными запасами сырой нефти и вывода его длины

Вывод результатов исполнения кода:

```
6825789460.784314
```

РЕШЕНИЕ В PANDAS

Количество стран в наборе данных можно вычислить с помощью применения функции len() к срезу данных, сформированному по выборке значений столбца «Страна». Код приведён в листинге 9.

```
1 len(df['Страна'])
```

Листинг 9 – Код для вычисления длины среза данных столбца «Страна»

Вывод результатов исполнения кода:

```
204
```

Количество стран в наборе данных, в которых подтверждённый объём запасов сырой нефти равен нулю, можно вычислить с помощью формирования среза данных с условием, как это показано в листинге 10.

```
1 len(df[df['Запасы сырой нефти'] == 0])
```

Листинг 10 – Код для вычисления длины среза данных с фильтрацией значений столбца «Запасы сырой нефти» по нулевому значению

Вывод результатов исполнения кода:

```
106
```

Среднее значение подтверждённых запасов сырой нефти можно получить, применив метод *mean()* к срезу набора данных по начениям столбца «Запасы сырой нефти». Код для вычисления среднего значения по столбцу приведён в листинге 11.

```
df['Запасы сырой нефти'].mean()
```

Листинг 11 – Код для вычисления длины среза данных с фильтрацией значений столбца «Запасы сырой нефти» по нулевому значению

Вывод результатов исполнения кода:

```
6825789460.784314
```

Сортировка среза данных по значениям столбца осуществляется с помощью вызова его метода *sort_values()*, с передачей методу параметра *by* со значением, соответствующим наименованию столбца, по которому необходимо произвести сортировку. Код для сортировки среза данных по столбцу «Страна» приведён в листинге 12. Результаты вывода представлены на рисунке 14.

```
df.sort_values(by=['Страна'])
```

Листинг 12 – Сортировка среза данных *df* по значениям столбца «Страна»

	Страна	Запасы сырой нефти
98	Afghanistan	0
57	Albania	199100000
15	Algeria	13420000000
199	American Samoa	0
14	Angola	13500000000
...	...	...
106	Virgin Islands	0
105	Western Sahara	0
29	Yemen	3160000000
104	Zambia	0
103	Zimbabwe	0

204 rows × 2 columns

Рис. 14 – Срез данных, отсортированный по значениям столбца «Страна», расположенным в алфавитном порядке

ЛАБОРАТОРНАЯ РАБОТА 3. РАБОТА С УСЛОВНЫМИ ОПЕРАТОРАМИ ПРИ ОБРАБОТКЕ ДАННЫХ

ЗАДАНИЕ

1. Определить страны с подтверждёнными запасами сырой нефти более 100 000 000 000.
2. Определить страны с подтверждёнными запасами сырой нефти в диапазоне от 10 000 000 000 до 30 000 000 000.

РЕШЕНИЕ В EXCEL

Значение ячейки в наборе данных можно получить с помощью функции INDEX(), которая в качестве первого аргумента при-

нимает набор ячеек и номер ячейки по порядку в данном наборе. Например, с помощью ввода следующей формулы возможно получить запасы сырой нефти Саудовской Аравии, так как она находится первой в списке стран, как это показано на рисунке 15:

```
=INDEX(provedoilreserves[Запасы сырой нефти];1)
```

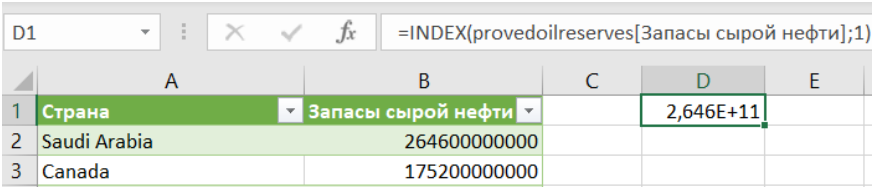
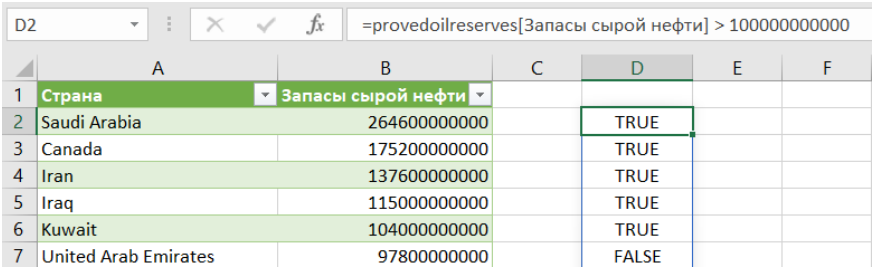


Рис. 15 – Получение в ячейке D1 значения подтверждённых запасов сырой нефти Саудовской Аравии.

Получить массив бинарных значений, показывающих больше ли значение ячейки какого-то определённого значения можно с помощью операции сравнения для ячейки. Для этого достаточно в ячейку 2 добавить следующую формулу:

```
=provedoilreserves[Запасы сырой нефти] > 100000000000
```

После ввода формулы в ячейку, она автоматически растянется на весь диапазон данных для сравнения. Ячейки в новом наборе примут значение TRUE, если запасы сырой нефти страны превышают 100 000 000 000, либо FALSE, если запасы не превышают этого значения. Полученный набор данных показан на рисунке 16.



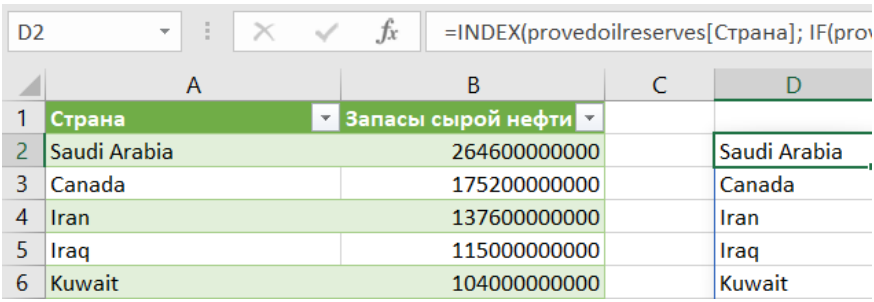
	A	B	C	D	E	F
1	Страна	Запасы сырой нефти				
2	Saudi Arabia	264600000000		TRUE		
3	Canada	175200000000		TRUE		
4	Iran	137600000000		TRUE		
5	Iraq	115000000000		TRUE		
6	Kuwait	104000000000		TRUE		
7	United Arab Emirates	97800000000		FALSE		

Рис. 16 – Определение ячеек, в которых значение больше 100 000 000 000.

Комбинируя вышеперечисленное с функцией ROW(), которая возвращает номер строки в диапазоне данных, для получения списка стран с запасами нефти более 100 000 000 000 можно использовать следующую формулу:

```
=INDEX(provedoilreserves[Страна];
IF(provedoilreserves[Запасы сырой нефти] >
100000000000; ROW(provedoilreserves[Страна]) - 1;
""))
```

ROW() – 1 используется так как первая строка в нашей таблице является заголовком. Список отфильтрованных стран приведён на рисунке 17.



	A	B	C	D
1	Страна	Запасы сырой нефти		
2	Saudi Arabia	264600000000		Saudi Arabia
3	Canada	175200000000		Canada
4	Iran	137600000000		Iran
5	Iraq	115000000000		Iraq
6	Kuwait	104000000000		Kuwait

Рис. 17 – Определение стран, в которых подтверждённые запасы сырой нефти больше 100 000 000 000.

Для получения списка стран, в которых подтверждённые запасы сырой нефти находятся в диапазоне от 10 000 000 000 до 30 000 000 000 в условный оператор можно добавить дополнительной условие с помощью логической функции AND(). Для этого в ячейку D2 необходимо ввести следующую формулу:

```
=INDEX(provedoilreserves[Страна];
IF(provedoilreserves[Запасы сырой нефти] >
10000000000;
MATCH(ROW(provedoilreserves[Страна]);ROW(provedoi
lreserves[Страна]));
""))*INDEX(provedoilreserves[Страна];
IF(provedoilreserves[Запасы сырой нефти] <
30000000000;
MATCH(ROW(provedoilreserves[Страна]);ROW(provedoi
lreserves[Страна])); ""))
```

Полученный список стран приведён на рисунке 18.

Страна	Запасы сырой нефти		
Saudi Arabia	264600000000		Kazakhstan
Canada	175200000000		Qatar
Iran	137600000000		China
Iraq	115000000000		United States
Kuwait	104000000000		Angola
United Arab Emirates	97800000000		Algeria
Venezuela	97770000000		Brazil
Russia	74200000000		Mexico

Рис. 18 – Определение стран, в которых подтверждённые запасы сырой нефти находится в диапазоне от 10 000 000 000 до 30 000 000 000.

РЕШЕНИЕ В PYTHON

Для получения списка стран с подтверждёнными запасами сырой нефти, превышающими 100 000 000 000 достаточно сформи-

ровать новых список, используя генератор списков с условием следующим образом:

```
filtered_values = [value for value in values if
  value['Запасы сырой нефти'] > 100000000000]
for value in filtered_values:
    print(value["Страна"])
```

Листинг 13 – Формирование списка стран с подтверждёнными запасами сырой нефти, превышающими 100 000 000 000

Вывод результатов исполнения кода:

```
Saudi Arabia
Canada
Iran
Iraq
Kuwait
```

Совмещение нескольких в условном операторе if возможно с помощью применения оператора and. Тогда вывести наименования стран, в которых подтверждённые запасы сырой нефти находятся в диапазоне от 10 000 000 000 до 30 000 000 000 можно следующим образом:

```
filtered_values = [value for value in values if
  value['Запасы сырой нефти'] >= 10000000000 and
  value['Запасы сырой нефти'] <= 30000000000]
for value in filtered_values:
    print(value["Страна"])
```

Листинг 14 – Формирование списка стран с подтверждёнными запасами сырой нефти, превышающими 100 000 000 000

Вывод результатов исполнения кода:

```
Kazakhstan
Qatar
```

```
China
United States
Angola
Algeria
Brazil
Mexico
```

РЕШЕНИЕ В PANDAS

Для решения данной задачи в pandas используется выборка из набора данных, которая позволяет сформировать срез набора данных на основе критериев, выражаемых логической функцией сравнения.

Срез данных, в котором значение ячеек столбца «Запасы сырой нефти» превышает значение 100 000 000 000 может быть получен следующим образом:

```
1 filtered_df = df[df["Запасы сырой нефти"] >
  100000000000]
2 filtered_df
```

Листинг 15 – Сортировка среза данных *df* по значениям столбца «Страна»

Результат выполнения кода представлен на рисунке 19.

	Страна	Запасы сырой нефти
0	Saudi Arabia	264600000000
1	Canada	175200000000
2	Iran	137600000000
3	Iraq	115000000000
4	Kuwait	104000000000

Рис. 19 – Определение стран, в которых подтверждённые запасы сырой нефти превышают 100 000 000 000.

При формировании среза данных в pandas возможно комбинировать несколько условий, например с помощью условного оператора &:

```
1 filtered_df = df[(df["Запасы сырой нефти"] >=
  10000000000) & (df["Запасы сырой нефти"] <=
  30000000000)]
2 filtered_df
```

Листинг 16 – Сортировка среза данных *df* по значениям столбца «Страна»

Результат выполнения данного кода представлен на рисунке 20.

	Страна	Запасы сырой нефти
10	Kazakhstan	30000000000
11	Qatar	25410000000
12	China	20350000000
13	United States	19120000000
14	Angola	13500000000
15	Algeria	13420000000
16	Brazil	13200000000
17	Mexico	12420000000

Рис. 20 – Формирование среза данных в pandas для стран с подтверждёнными запасами сырой нефти, находящимися в диапазоне от 10 000 000 000 до 30 000 000 000

ЗАКЛЮЧЕНИЕ

Выполнение лабораторных работ из данных методических указаний рекомендуется осуществлять постоянно экспериментируя с аргументами функций и по возможности реализуя дополнительные способы обработки данных, которые не описаны в методических указаниях, однако доступны в среде редактирования кода в виде всплывающих подсказок и встроенной документации.

При выполнении лабораторных работ не следует просто переписывать код из примеров, вместо этого следует шаг за шагом разбираться со значением каждой строчки кода, пользуясь вспомогательной документацией при необходимости.

Отдельно следует отмечать различие в подходах между представлением данных в виде стандартных коллекций Python – списков, словарей, кортежей; и наборами данных в pandas, где используется принципиально другая логика, заключающаяся в формировании срезов данных.

Оформление отчёта по лабораторной работе должно соответствовать действующим государственным стандартам. К отчётам по лабораторным работам должны прилагаться исходные файлы Excel, Python и pandas. В случае использования Jupyter Notebook достаточно приложить файлы *.ipynb.

В работе содержится только информация, непосредственно связанная с этапами работы. Все иллюстрации раскрывают содержание выполненных действий. При написании отчёта не следует заимствовать фрагменты из учебников и интернета. Достаточно в тексте оформить ссылки на них.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. У. Маккини.Python и анализ данных. – ДМК-Пресс, 2020. – 540 с.
2. Леора С.Н., Бурнаева Э.Г. Обработка и представление данных в MS Excel. Учебное пособие. – Лань, 2018. – 156 с.
Подробнее: https://www.labirint.ru/books/516014/. – 2010. – 432 с.
3. Python documentation [Электронный ресурс]. – URL: https://docs.python.org/3/ (дата обращения 10.02.2020).
4. Pandas documentation [Электронный ресурс]. – URL: https://pandas.pydata.org/docs/ (дата обращения 10.02.2020).
5. ГОСТ 7.32-2017 Система стандартов по информации, библиотечному и издательскому делу. Отчет о научно-исследовательской работе. Структура и правила оформления. – М.: Стандартинформ, 2018. – 33 с.
6. ГОСТ 2.105-95. Единая система конструкторской документации. Общие требования к текстовым документам. – М.: Стандартинформ, 2007. – 32 с.
7. ГОСТ Р 7.0.5-2008. Библиографическая ссылка. Общие требования и правила составления. – М.: Стандартинформ, 2008. – 23 с.

СОДЕРЖАНИЕ

ИНФОРМАТИКА. СТАТИСТИЧЕСКАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Методические указания по выполнению лабораторных работ для студентов бакалавриата направления 21.03.01

Сост.: *С.Б. Крыльцов, М.А. Коробицына*

Печатается с оригинал-макета, подготовленного кафедрой информатики и компьютерных технологий

Ответственный за выпуск *С.Б. Крыльцов*

Лицензия ИД № 06517 от 09.01.2002

Подписано к печати 14.04.2021. Формат 60×84/16.
Усл. печ. л. 1,9. Усл.кр.-отт. 1,9. Уч.-изд.л. 1,5. Тираж 75 экз. Заказ 315.

Санкт-Петербургский горный университет
РИЦ Санкт-Петербургского горного университета
Адрес университета и РИЦ: 199106 Санкт-Петербург, 21-я линия, 2