УДК 657.1:004

СОЗДАНИЕ ОТЧЁТОВ ПО ДОХОДАМ ОРГАНИЗАЦИЙ С ПОМОЩЬЮ ЯЗЫКА PYTHON

Ильичев В.Ю.,

к.т.н., доцент,

Калужский филиал ФГОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»,

Калуга, Россия

Юрик Е.А.,

к.т.н., доцент,

Калужский филиал ФГОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»,

Калуга, Россия

Аннотация

В статье описано исследование, целью которого являлась разработка компьютерного приложения для обработки баз данных, содержащих экономические показатели организаций, с использованием современного языка программирования Python и подключённых библиотек: Pandas, предназначенной для работы с базами данных и Matplotlib, предназначенной для вывода качественной графической информации.

Произведён обзор существующих на данный момент публикаций по данной тематике, размещённых в сети интернет и в документации к вышеперечисленным программным продуктам.

Разработана и описана последовательность команд, необходимых для подключения библиотек, загрузки данных из файла и создания необходимой

исследователю выборки. Показан процесс создания отчёта по результатам сортировки полученной выборки с помощью команд языка Python.

Описанная последовательность действий продемонстрирована на примере обработки массива данных по экономическим показателям (доходам, расходам и др.) организаций Ленинградской области за 2017-2018 гг. Массив отсортирован по доходам за 2017 г. и выбраны компании с наибольшим значением данного экономического показателя. Для наглядного сравнения и анализа результаты обработанной выборки из массива представлены в виде совместной диаграммы доходов организаций за два года с использованием библиотеки Matplotlib.

Сделаны выводы о преимуществах описанного подхода к обработке экономических данных по сравнению с традиционным применением существующих программных комплексов. Он отличается простотой использования и подходит для обработки экономических данных, представленных практически в любом формате, в том числе малораспространённом.

Даны рекомендации по дальнейшему развитию разработанной методики в областях наиболее передовых экономических исследований.

Ключевые слова: экономические показатели, доход организации, база данных, программа, язык Python, библиотека Pandas, библиотека Matplotlib.

CREATING ORGANIZATION REVENUE REPORTS USING PYTHON

Ilichev V.Y.,
PhD, Associate Professor,
Kaluga Branch of Bauman Moscow State Technical University,
Kaluga, Russia

Yurik E.A.,
PhD, Associate Professor,
Kaluga Branch of Bauman Moscow State Technical University,
Kaluga, Russia

Annotation

The article describes a study aimed at developing a computer application for processing databases containing economic indicators of organizations, using a modern Python programming language with the connection of libraries: Pandas designed to work with databases and Matplotlib designed to output high-quality graphical information.

An overview of the current publications on this subject, posted on the Internet and in the documentation for the above-mentioned software products, is made.

The sequence of commands required to connect libraries, load data from a file, and create the selection required by the researcher has been developed and described. The following shows how to generate a report based on the results of sorting the obtained sample using Python commands.

The described sequence of actions is demonstrated on the example of processing of the array of data on economic indicators (income, expenses, etc.) of the organizations of Leningrad region for 2017-2018 years. The array is sorted by income for 2017 year and 25 companies with the largest value of this economic indicator are selected. For visual comparison and analysis, the results of the processed sample from the array are presented in the form of a joint diagram of the income of the organizations for two years using the Matplotlib library.

Conclusions were drawn on the advantages of the described approach to economic data processing compared to the traditional application of existing software complexes. It is easy to use and is suitable for processing economic data presented in almost any format, including low-volume data.

Recommendations are given for further development of the methodology in the fields of the most advanced economic research.

Keywords: economic indicators, organization income, database, program, Python language, Pandas library, Matplotlib library.

Введение. Решение экономических задач, как научных, так и прикладных, при современном уровне развития промышленности, торговли и других сфер жизнедеятельности людей, не представляется возможным без использования компьютерной техники и технологий программирования [5]. Особенно стоит выделить такой, пожалуй основной, аспект экономических исследований как обработка огромных массивов статистических данных [7]. Эти данные могут быть представлены в разных форматах файлов – как распространённых (Excel, 1C, Access, Statistica), так и редких (данные отчётов по работе технологического оборудования).

Такие исследования в англоязычной литературе называют Data Science [1] – «наукой работы с данными», включающей в себя изучение процедур обработки данных, машинного обучения специалистов и визуализации данных [9].

В экономике в качестве данных для анализа обычно используются объёмы производства, продаж и доходов различных компаний и организаций. Получаемые массивы информации условно можно представить в виде столбцов численных и литерных данных с не отсортированными по какому-либо признаку строками [8].

Для формирования и обработки таких массивов традиционно используют СУБД (системы управления базами данных), из которых для выполнения экономических задач в настоящее время можно выделить, например, продукты PostgreSQL, Clickhouse, Arenadata DB и др. [10]

Освоение современных СУБД является сложным и длительным процессом, тем более что каждая из них отличается своими особенностями и предпочтительными сферами применения. Поэтому для решения сравнительно несложных задач в сфере экономики, таких как выбор из массива необходимых данных и создание по ним наглядных отчётов авторами предлагается один из популярнейших универсальных языков программирования Python [3]. Для выполнения специфических функциональных задач, таких как например обработка баз данных или вывод графиков, существуют отдельно подключаемые библиотеки, в свою очередь также написанные на языке Python. При использовании данного подхода получается, что всё приложение, выполняющее последовательно несколько задач, создаётся с помощью одного языка программирования и достаточно несложных для понимания даже рядового пользователя команд.

Цель исследования. Характерный пример сайта, на котором можно получить много статистической информации для научных исследований в экономической отрасли – сайт Федеральной налоговой службы России, а именно открытые данные, выложенные в соответствии с Налоговым кодексом - информация о деятельности государственных органов и органов местного самоуправления, размещенная в сети «Интернет» в виде массивов данных в форматах, обеспечивающем их автоматизированную обработку, и на условиях её свободного (бесплатного) использования [12].

Среди выложенных файлов данных многие отличаются большим объёмом (до нескольких гигабайт) и естественно вручную обработать их невозможно. Кроме того, данные представлены в разных форматах – csv, xlsx (Excel), Parquet, HDF5 и многих других. Традиционно считается, что для каждого формата необходима предназначенная для его загрузки и обработки программа и(или) конвертер (который должен корректно работать). При этом абсолютное большинство программных продуктов для работы с базами данных являются коммерческими и очень дорогими.

В настоящее время появилась возможность использования альтернативного решения – применения свободно распространяемого языка Python, для которого существует библиотека обработки файлов данных Pandas, обладающая широкими возможностями и постоянно совершенствующаяся.

Целью данного исследования является разработка последовательности создания программы для обработки данных по отчётным экономическим показателям организаций и создания качественного графического материала по результатам обработки. Указанный процесс проиллюстрирован примером.

Материал и методы исследования. Описание последовательности выполнения команд разрабатываемой программы частично можно обнаружить в документации к языку Python [4] и к библиотеке Pandas [15], однако данной информации оказывается недостаточно для практического использования. Поэтому пришлось изучить множество форумов по программированию и имеющуюся немногочисленную пока литературу, например, [11, 13].

В результате была разработана программа, включающая следующие блоки:

1. подключение необходимых библиотек команд;

2. чтение файла-массива данных и автоматическая конвертация его в формат Pandas;

3. выбор в массиве данных столбцов (полей), содержащих только необходимую исследователю информацию;

4. выделение поля-признака, по значениям которого будет осуществляться сортировка всех строк (по возрастанию или убыванию этих значений);

5. изменение названий выбранных полей на удобные и понятные для отображения информации;

6. определение необходимого для дальнейшего анализа количества первых отсортированных по полю-признаку объектов (выборки);

7. команды для формирования графика: выбор типа диаграммы, названий осей, числовых меток, задание цветов, форматирование всех элементов для наиболее

качественного отображения;

8. построение графика по сформированной в п. 6 выборке.

Для отработки указанной последовательности действий программы использован файл, содержащий информацию о суммах доходов и расходов по данным бухгалтерской (финансовой) отчетности организаций (ООО, АО и других форм собственности) Ленинградской области за 2017-2018 гг. с сайта [2]. Файл представлен в формате csv и состоит из 11 столбцов (полей) и 17884 строк, первая из которых содержит наименование поля, остальные – соответствующие данные по организациям (табл. 1).

Таблица 1 - Описание полей базы данных «Сведения о суммах доходов и расходов организаций»

№ поля	Наименование поля	Описание поля
1	Id	Идентификатор строки
2	Name2017	Наименование и тип организации из данных 2017-го года (напр., ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "СИГМА", АКЦИОНЕРНОЕ ОБЩЕСТВО "ИСКРА-ПРИБОР")
3	Name2018	Наименование и тип организации из данных 2018-го года. Если поля Name2017 и Name2018 не пустые, то значит организация изменила название в 2018 году.
4	INN	ИНН организации
5	Income2017	Сума доходов за 2017 год, руб.
6	Expenses2017	Сумма расходов за 2017 год, руб.
7	Balance2017	Разница между доходами и расходами в 2017 году (поле отсутствует в исходной БД и добавлено для удобства просмотра)
8	Income2018	Сумма доходов за 2018 год, руб.
9	Expenses2018	Сумма расходов за 2018 год, руб.
10	Balance2018	Разница между доходами и расходами в 2018 году (поле отсутствует в исходной БД и добавлено для удобства просмотра)
11	Region	Регион, в котором был получен ИНН

Допустим, что перед исследователем поставлена задача выбрать несколько организаций с наибольшим доходом за 2017 г. и вывести данные об их сравнительных доходах за 2017 и 2018 гг. в графическом виде.

Следовательно, из всех полей для построения графика необходимо выделить только три - №2 (Name2017), содержащее наименование и форму собственности организации, №5 (Income2017), содержащее сумму доходов за 2017 г. и №8 (Income2018), содержащее сумму доходов за 2018 г.

Затем для наглядности нужно построить гистограммы по доходам за два указанных года на одном графике. Все средства для выполнения указанных задач имеются в языке Python с подключёнными специальными библиотеками.

Первые несколько так называемых кадров данных (совокупности строк) частично представлены на рис. 1 для лучшего понимания организации исследуемой базы данных.

Id	Name2017	Name2018	INN	Income2017	Expenses2017	Balance2017	Income2018	Expenses2018	Balance2018
276588	ОТКРЫТОЕ А		###	225168000	216819000	8349000	217580000	214099000	3481000
1416178	ЗАКРЫТОЕ А	АКЦИОНЕР	###	1351882000	1171329000	180553000	2163343000	1765222000	398121000
1798572		АКЦИОНЕР	###				4342286000	3254336000	1087950000
1063795	ОТКРЫТОЕ А		###	61451000	62329000	-878000	17584000	19874000	-2290000
1433449	ОТКРЫТОЕ А		###	241369000	228812000	12557000	148686000	101080000	47606000
1785333	ЗАКРЫТОЕ А		###	35916000	36581000	-665000	30694000	31903000	-1209000
1818636	АКЦИОНЕРН		###	160464000	162701000	-2237000	329895000	329342000	553000
495336		АКЦИОНЕР	###				4335762000	3123307000	1212455000
1991508	ЗАКРЫТОЕ А		###	102500000	97932000	4568000	42562000	42245000	317000
1538581	ЗАКРЫТОЕ А		###	0	182000	-182000	0	301000	-301000
1677597	АКЦИОНЕРН		###	11344000	15523000	-4179000	11371000	10751000	620000
1677511	ОБЩЕСТВО		###	183000	189000	-6000	36000	56000	-20000
1538320	ОБЩЕСТВО		###	1040000	950000	90000	1808000	1722000	86000
1962779	ОБЩЕСТВО		###	1760000	1665000	95000	1736000	1647000	89000
1990798	ОБЩЕСТВО		###	0	0	0	0	0	0

Рис. 1 - Кадры данных исследуемого массива

Далее рассмотрим результат выполнения вышеуказанных блоков команд в отношении выбранного массива данных.

При выполнении примера последовательно решались следующие задачи:

1. Сделана выборка по 25 организациям, имеющим наибольший суммарный доход за 2017 г. с формированием массива по данным доходам и его сортировкой по убыванию.

2. Создана выборка по суммарным доходам тех же организаций за 2018 г.

3. Полученные значения суммарных доходов выведены по выбранным организациям в виде сравнительной диаграммы.

Результаты. Результат выполнения программы приведён на рис. 2. Вывод и оформление этой диаграммы осуществлён с использованием специальной библиотеки для работы с графическими данными Matplotlib для Python.

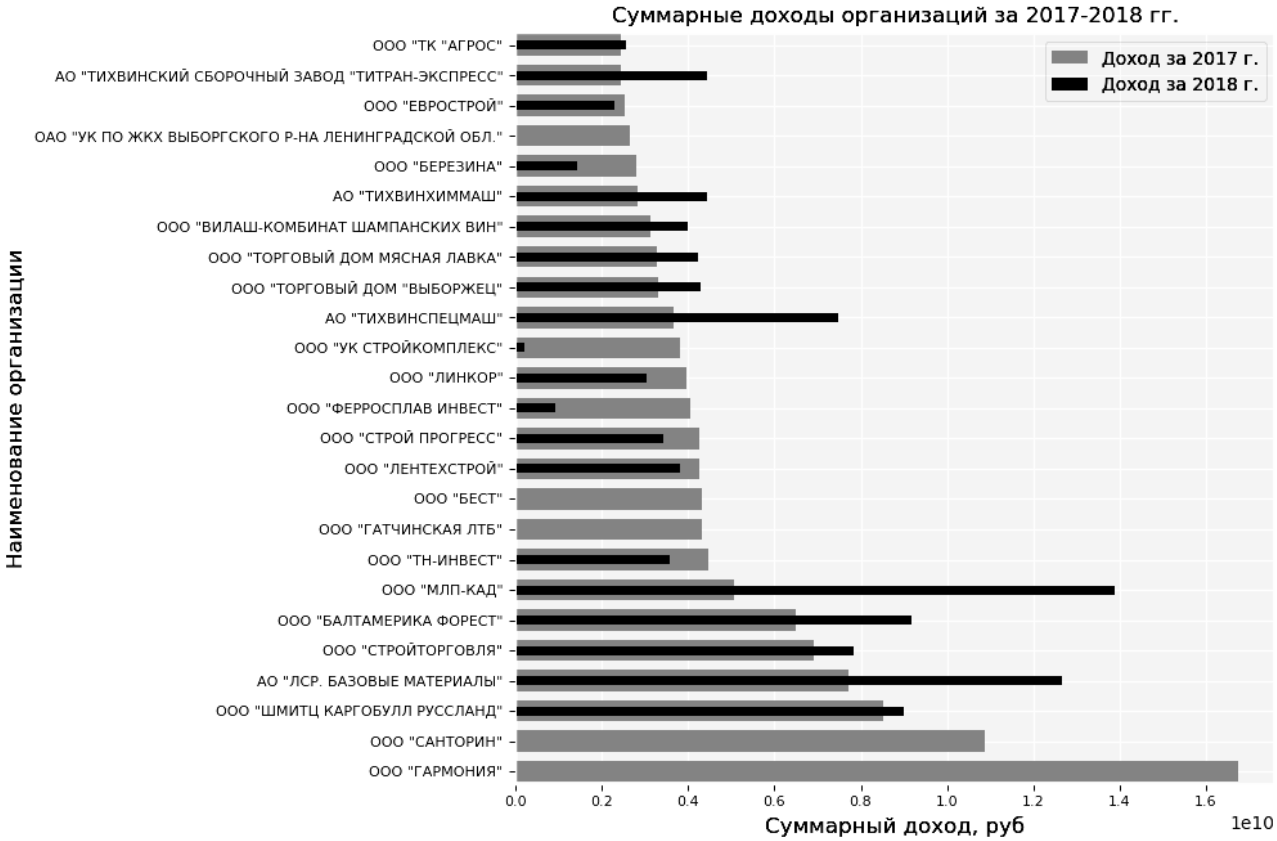

Рис. 2 - Графический результат обработки массива данных

По выведенному на экран (и автоматически записанному в файл) графику можно наглядно проанализировать изменение суммарных доходов организаций, имеющих максимальное значение этого показателя по итогам

2017 г. Видно, на сколько у некоторых организаций он вырос, и на сколько сократился в той или иной мере у других. Для некоторых организаций, прекративших своё существование, либо сменивших форму собственности, значения доходов за 2018 г. в базе данных и соответственно на рис. 2 не представлены.

Графики, оформленные подобным или иным образом с помощью библиотеки Matplotlib можно использовать при подготовке докладов и презентаций, для размещения в сети интернет, для дальнейшего экономического анализа (например, выявления тренда и прочих закономерностей). В данной библиотеке предусмотрены очень широкие возможности создания любого графического оформления результатов исследований.

С использованием дополнительно подключаемых к программе на языке Python библиотек можно также оформить пользовательский интерфейс (GUI), реализовать различные виды анализа статистических данных (например, с помощью нейросетевых моделей), выводить результаты в виде таблиц, трёхмерных моделей и многое другое.

Обсуждение. Данное исследование было посвящено лишь одному аспекту применения языка Python для исследований в области экономики. Необходимо отметить, что именно эта область отличается от других созданием и хранением огромных массивов данных, объём которых с течением времени увеличивается в геометрической прогрессии. Это связано с непрерывным техническим и социальным прогрессом общества и экономик стран мира. При этом всё большему количеству экономистов приходится при подготовке различного рода отчётов пользоваться программированием для автоматизации своего труда. Между тем, неспециалисту в области информатики сложно освоить существующие традиционные «тяжёлые» программные продукты. Этим и объясняется бурное развитие языка программирования широкого

профиля Python, характеризующегося простотой освоения и удобством использования для разработки прикладных программ любого назначения.

Приведённое в данной статье исследование также направлено на решение прикладной задачи в области экономики, и иллюстрирует процесс и результат использования языка Python с подключёнными библиотеками Pandas и Matplotlib.

Заключение. По результатам проведённой работы можно сделать следующие выводы:

- данный подход отличается быстрой загрузкой и обработкой данных практически любого формата, к тому же имеющих пропуски и другие особенности;
- процесс разработки программы статистического экономического анализа, начиная с загрузки массива данных и заканчивая выводом результатов, в данном случае осуществляется в одном крайне простом приложении, практически в текстовом редакторе;
- доказана простота создания любых необходимых выборок данных из массива;
- вывод результатов осуществляется в виде графического материала полиграфического качества;
- команды языка Python отличаются простотой освоения и использования.

Эти и другие достоинства рассмотренного подхода к анализу данных позволяет рекомендовать его к использованию в различных современных отраслях экономики: системной статистике [6], исследовании закономерностей экономического развития, управлении предприятиями, моделировании пространственной структуры развития экономики и прочих [14]. Авторами данной статьи планируется использование языка Python с библиотеками, написанными на этом же языке, для формирования моделей экономического развития предприятий крупнейшего инновационного региона - Калужской области.

Библиографический список:

1. Борисов Г.Б., Краснов М.Р. Анализ макроэкономических показателей в мировой экономике методами Data Science. В сборнике: XIV Неделя науки молодежи СВАО посвященная 85-летию со дня рождения Ю.А. Гагарина. Сборник статей по итогам работы научных конференций и круглых столов в рамках XIV Недели науки молодежи северо-восточного административного округа города Москвы. - 2019. - С. 22-24.

2. Данные для бизнеса. Сведения о суммах доходов и расходов по данным бухгалтерской (финансовой) отчетности организаций (ООО, АО). [Электронный ресурс]. - Режим доступа — URL: https://mydata.biz/ru/catalog/databases/firm_incomes (дата обращения: 03.06.2020).

3. Дж. Вандер Плас. Python для сложных задач: наука о данных и машинное обучение. - Санкт-Петербург: Питер, 2018. - 572 с.

4. Документация по библиотекам Python с примерами. [Электронный ресурс]. - Режим доступа — URL: https://pythonru.com/biblioteki (дата обращения: 03.06.2020).

5. Дружиловская Э.С. Современные проблемы бухгалтерского учета и отчетности с точки зрения аналитиков / Э.С. Дружиловская // Международный бухгалтерский учет. – 2015. – № 10. – С. 54-64.

6. Ильичев В.Ю., Чухраев И. В., Юрик Е.А. Применение методов компьютерного статистического анализа для прогнозирования потребления электрической энергии // Информационно-измерительные и управляющие системы. - 2020. - Т. 18. - № 2. - С. 24-32.

7. Сидорова М. И. Современные информационные технологии как инструмент автоматизации бухгалтерского учета и отчетности / М. И. Сидорова // Международный бухгалтерский учет. – 2011. – № 28. – С. 19-24.

8. Слободняк И. А., Пискунов И. В. Актуальные проблемы автоматизации бухгалтерского учета и отчетности / И. А. Слободняк, И. В. Пискунов //

Бухгалтерский учет в бюджетных и некоммерческих организациях. – 2014. – № 7. – С. 29-34.

9. Сорокина Л. Н. Проблемы внедрения автоматизации учета и подготовки отчетности в условиях перехода на международную систему финансовой отчетности / Л. Н. Сорокина // Финансовая аналитика: проблемы и решения. – 2014. – № 3. – С. 13- 17.

10. СУБД для цифровой экономики. М.: Открытые системы. СУБД. - 2018. - № 1. - С. 3-9.

11. Сысоева М.В., Сысоев И.В. Программирование для «нормальных» с нуля на языке Python. Учебник. В двух частях. Часть 1. - М.: ООО «МАКС Пресс». 2018. - 176 с.

12. Федеральная налоговая служба. Открытые данные. [Электронный ресурс]. — Режим доступа — URL: https://www.nalog.ru/opendata/ (Дата обращения 03.06.2020)

13. Хайбрахманов С.А. Основы научных расчётов на языке программирования Python: учебное пособие. - Челябинск: Изд-во Челябинского государственного. университета, 2019. - 96 с.

14. Шитова Т. Ф. Использование передовых информационных технологий в бухгалтерском учете / Т. Ф. Шитова // Международный бухгалтерский учет. – 2012. – № 22. – С. 21-26.

15. Pandas documentation. Date: Mar 18, 2020. Version: 1.0.3. [Электронный ресурс]. - Режим доступа — URL: https://pandas.pydata.org/docs/pandas.pdf (дата обращения: 03.06.2020).

Оригинальность 90%