Анализ данных в Python
модуль 1, учебный год 2024–2025

Сергей Головань
Российская экономическая школа
sgolovan@nes.ru

Учебный ассистент: Геннадий Иванов (givanov@nes.ru)

Описание курса

Курс «Анализ данных в Python» является введением в статистический анализ данных с использованием программного обеспечения с открытым кодом из экосистемы языка Python. В настоящее время накоплены огромные массивы данных. В то же время для их обработки, анализа и визуализации часто используются либо устаревшие, либо неоправданно дорогие программные средства. Сейчас все чаще для такого анализа используется открытое программное обеспечение, в частности язык Python и его богатый набор библиотек. Целью этого курса является обучение студентов инструментам, позволяющим эффективно обрабатывать большие объемы данных, описывать и визуализировать их, а также использовать для принятия экономических решений. Курс учит, как очистить данные от ошибок, как агрегировать данные, исследовать и представить результаты анализа, используя современные графические возможности библиотек языка Python.

Этот курс является обязательным. Он состоит из 14 лекций и 7 семинаров.

Система оценивания и требования к выставлению итоговой оценки

Курс не требует никаких предварительных знаний, кроме базовых курсов по теории вероятностей.

В курсе будут предложены 5 домашних заданий, которые составят 50% от окончательной оценки за курс. Остальные 50% приходятся на финальный экзамен, который будет задан на дом.

Содержание курса

1. Организация программного кода и данных
   (a) Введение в системы управления версиями
   (b) Хранение и организация данных
2. Основы языка Python

   (a) Типы данных: списки, кортежи, словари, строки, числа, булевый тип
   (b) Генераторы списков
   (c) Условные операторы, циклы
   (d) Функции, классы
   (e) Повторное использование кода

3. Обработка данных
   (a) Загрузка данных из сети, csv, Excel
   (b) Структуры данных: Series, DataFrame, Panel
   (c) Объединение нескольких наборов данных
   (d) Индексирование данных, выборки
   (e) Вычислительные средства для модификации данных
   (f) Группировка и агрегирование данных
   (g) Преобразование формы данных
   (h) Представление даты и времени

4. Визуализация данных
   (a) Создание простых диаграмм: график, диаграмма рассеивания, столбчатая диаграмма
   (b) Вывод графика из нескольких источников
   (c) Совершенствование внешнего вида диаграмм
   (d) Дополнительные библиотеки для интерактивной работы с графикой
   (e) Распределенная визуализация

Структура и организация учебной дисциплины

Лекции будут следовать от мотивационных примеров и примеров экономических моделей к общим утверждениям и принципам. Кроме того, студентам будут выданы компьютерные упражнения, которые позволят освоить методы анализа и визуализации данных на практике.

Примеры заданий и вопросов для самостоятельной работы и промежуточного контроля

1. Источник данных: https://www.seattle.gov/transportation/projects-and-programs/programs/new-mobility-program/scooter-bike-share-data

   Ссылка на данные: https://s3.amazonaws.com/pronto-data/open_data_year_two.zip

   Пример анализа данного набора данных: https://jakevdp.github.io/blog/2015/10/17/analyzing-pronto-cycleshare-data-with-python-and-pandas/

   (a) Подключите необходимые пакеты.

(b) Покажите содержимое архивного файла `cycle_share.zip`.

(c) Импортируйте `2016_trip_data.csv` непосредственно из архива в Pandas DataFrame. При этом тип переменных *starttime* и *stoptime* должен быть `datetime64[ns]`, а переменные *usertype* и *gender* должны быть категоризованными. Выведите на печать типы всех переменных. Выведите на печать пять строк для следующих колонок: *trip_id*, *Date*, *starttime*, *tripduration*, *gender*. Выведите на печать число наблюдений в каждой категории для двух категоризованных переменных в данном наборе. Сохраните данные в файл `exam_data.hdf`, таблица *trips*.

(d) Импортируйте `2016_weather_data.csv` непосредственно из архива в Pandas DataFrame. При этом тип переменной *Date* должен получиться `datetime64[ns]`, а переменная *Events* должна быть категоризованной. Выведите на печать типы всех переменных. Выведите на печать пять строк для следующих колонок: *trip_id*, *Date*, *starttime*, *tripduration*, *gender*. Заметьте, что значения в колонке *Events* могут быть скажем, «Fog-Rain» и «Fog , Rain». Унифицируйте эти значения. Выведите на печать число наблюдений в каждой категории для колонки *Events*.

(e) Найдите наиболее популярные пары станций. Заметьте, что поездка из *A* в *B* считается так же, как и поездка из *B* в *A*. Также исключите поездки, которые начинаются и заканчиваются на одной и той же станции. Сохраните результаты в файле Excel. Выведите на печать названия двух наиболее популярных станций назначения из списка.

(f) Нарисуйте плотность логарифма длительности поездки.

(g) Нарисуйте плотность логарифма длительности поездки для мужчин и женщин отдельно. Исключите наблюдения со значениями «Other» в колонке *gender*.

(h) Сосчитайте число поездок на каждую дату и для каждого типа пользователя. Изобразите графики двух временных рядов (для каждого типа пользователя) на одном графике. Так как оба они обладают высокой сезонностью, агрегируйте данные до недельных.

(i) Вычислите среднюю продолжительность поездки для каждого типа осадков (колонка *Events* в наборе данных о погоде).

(j) Наблюдается ли зависимость (раздельно для каждого пола) между продолжительностью поездки (в логарифмах) и следующими переменными, отвечающими за погоду: средняя температура в градусах Цельсия и логарифм объема осадков? Изобразите две диаграммы рассеивания и проведите соответствующие регрессионные прямые для каждой объясняющей переменной для каждого пола в отдельности. Исключите наблюдения со значениями «Other» в колонке *gender*. *Указание*: $T_F = 1.8 \cdot T_C + 32$.

Политика академической честности

Списывание, плагиат и другие нарушения академической этики в РЭШ недопустимы.