Python for Economics

Zhenhua He
Richard Lawrence

Morning session, 10/1/2021

Table of Contents

This course is divided into numbered lessons

17. Web Scraping with BeautifulSoup

- Advanced exercise: Candlestick plot
- Wrap-up

Lesson 17
Web Scraping with BeautifulSoup

Use Python libraries to extract data from a website

Applying Lesson Learning Objectives

After this lesson, you will know:

- What are regular expressions
- What is BeautifulSoup library and how to use it

During this lesson, you will use:

- What is web scraping
- HTML Basics
- What is requests library and how to use it

What is web scraping

- **Scraping**: a process of extracting (from various means), copying and screening of data.
- **Web Scraping**: scraping or extracting data or feeds from the web (like from web-pages or websites)

Why does one scrape the web?

- Collect data for Research
- Sales and Marketing
- Products prices & popularity comparison
- …

Regular Expression

- A sequence of characters forms a search **pattern**
- `findall()`: Check if a string contains the pattern
- Python built-in module **re**
- Example

```
import re
text = "I went to TAMU at 8 AM"
pattern = "AM"
result = re.findall(pattern, text)
["AM", "AM"]
```

Regular Expression

re.compile()

- convert a pattern string into a pattern **object**
- **Object** combines the pattern and `findall()` function
- Example

```
text = "I went to TAMU at 8 AM"
pattern = re.compile("AM")
result = pattern.findall(text)
```

- No need to enter the pattern string again

Regular Expression Examples and Exercises

Go to Google Classroom Lesson 17 “Regular Expressions”

Tasks

- Follow instructions for the examples
- Work on the exercises (**due** by 10/1 6:00 PM)

BeautifulSoup

- A Python library for pulling data out of HTML files.
- Works with a *parser* to provide idiomatic ways of navigating, searching, and modifying the source tree.
- Built-in parser available (we will use)

Review: JSON - JavaScript Object Notation

- A text format for storing data
- language-independent
- The data to extract could be in JSON format.

JSON string examples:

```
'{"name":"Jack", "age":20, "major":"computer science"}'
```

```
'{"args": {}, "data": "", "files": {},
  "form": { "soup": "hot soup"}, ... }'
```

from the Requests exercise we did previously

JSON module

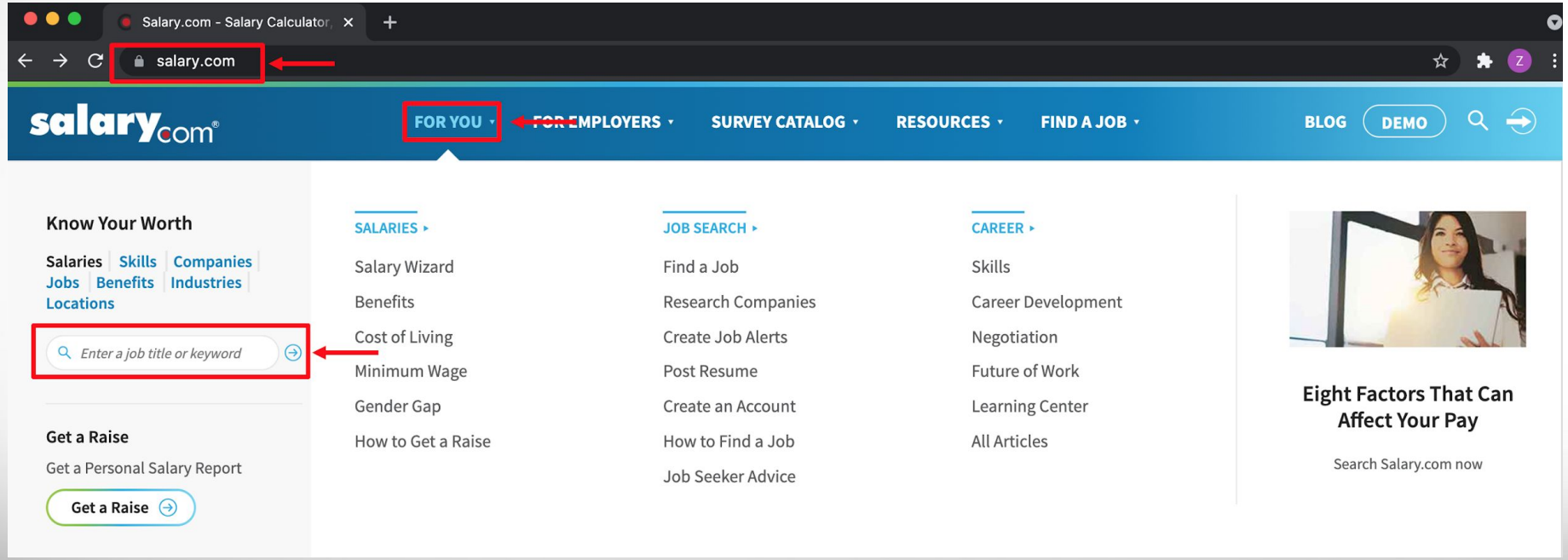
- Python built-in module **json**
- `json.loads()`: converts JSON string to Python dictionary
- Example

```
import json
text = "{ keys : values, … }"
dict = json.loads(text)
```

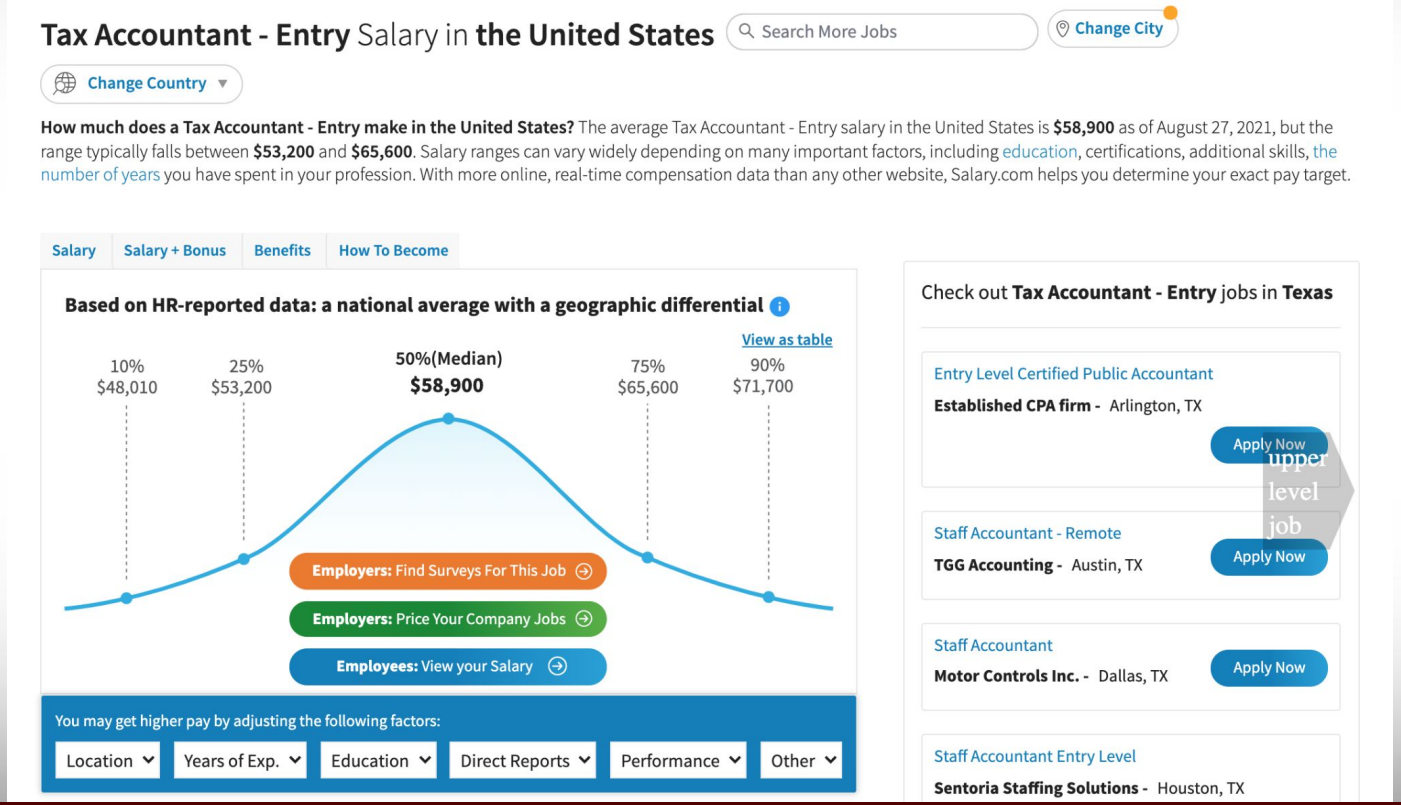
Today's Exercise: Scraping a website

- Collect salary stats from https://www.salary.com/
- Salary stats for a single city
- Salary stats for multiple cities
- Save data in a file

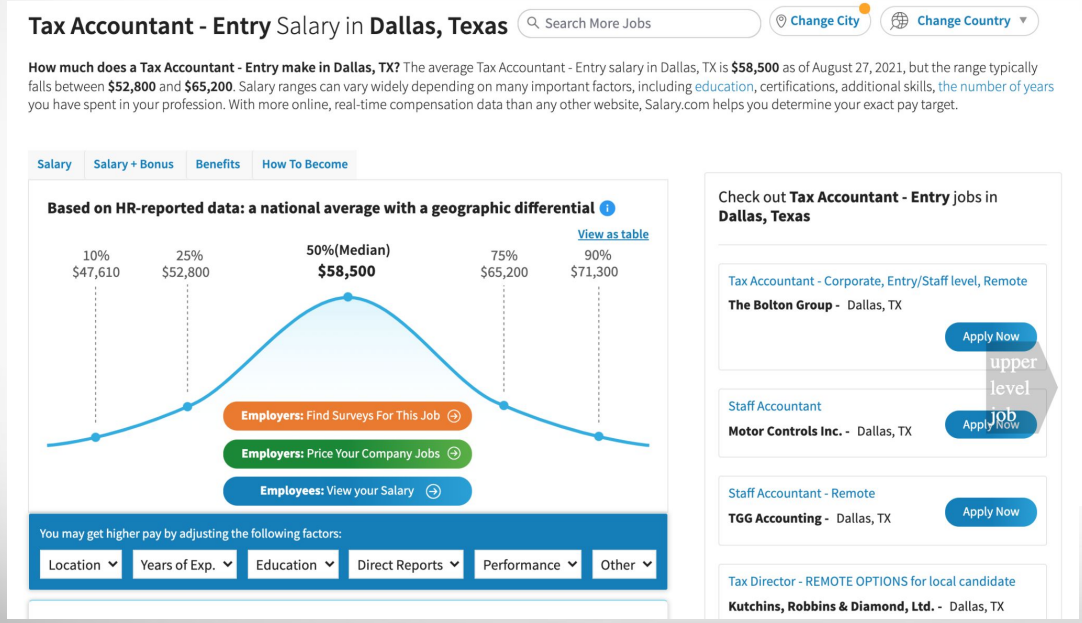
Investigate salary stats

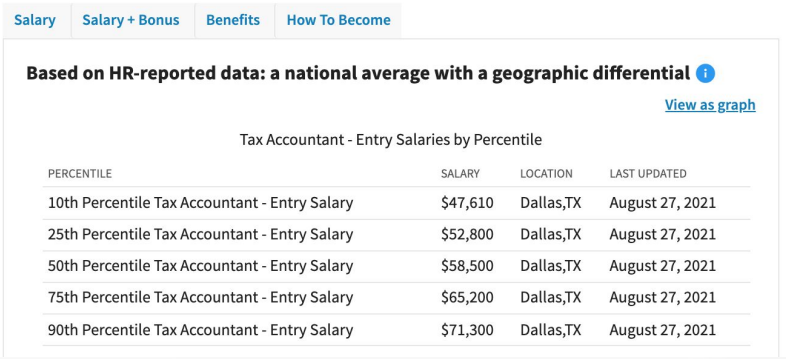


Analyze the salary stats



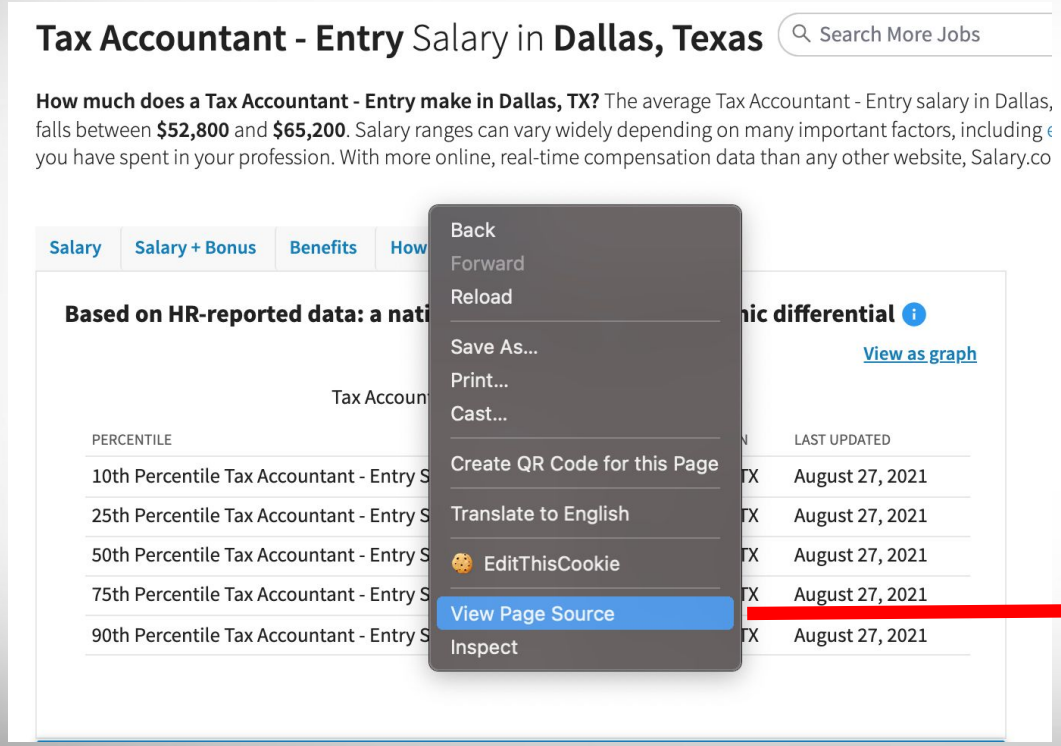
Collect the salary stats for Dallas, TX





upper level job

Check the page source




```
<!-- Occupation -->

<script type="application/ld+json">
{
    "@context": "http://schema.org",
    "@type": "Occupation",
    "name": "Tax Accountant - Entry",
    "mainEntityOfPage": {
        "@type": "WebPage",
        "lastReviewed": "2021-08-27T00:00:00Z"
    },
    "description": "The Tax Accountant - Entry helps with
    "estimatedSalary": [{
    "@type": "MonetaryAmountDistribution",
    "name": "base",
    "currency": "USD",
    "duration" : "P1Y",
    "percentile10": "47610",
    "percentile25": "52800",
    "median": "58500",
    "percentile75": "65200",
    "percentile90": "71300"
    },
    {
    "@type": "MonetaryAmountDistribution",
    "name": "total compensation",
    "currency": "USD",
    "duration" : "P1Y",
    "percentile10": "48473",
    "percentile25": "54300",
    "median": "60700",
    "percentile75": "67800",
    "percentile90": "74264"
    }]

        ,
        "occupationLocation": [{
            "@type": "City",
            "name": "Dallas, TX"
        }]

}
</script>
```

Collect stats for multiple cities – structured URL

Senior accountant Dallas, Texas	https://www.salary.com/research/salary/alternate/senior-accountant-salary/dallas-tx
Entry data analyst Miami, Florida	https://www.salary.com/research/salary/alternate/entry-data-analyst-salary/miami-fl
Tax accountant entry Washington, DC	https://www.salary.com/research/salary/alternate/tax-accountant-entry-salary/washington-dc

```
url_template = 'https://www.salary.com/research/salary/alternate/{}-salary/{}'
```

Goal: CSV file with collected data

salary_data_multi_city.csv

1 to 5 of 5 entries Filter

Title	Location	Description	10%	25%	50%	75%	90%
Tax Accountant - Entry	New York, NY	The Tax Accountant - Entry helps with the preparation of local, state and federal level returns to be submitted within specified tax deadlines. Assists in the maintenance and preparation of tax-related items to include tax records, tax returns, tax schedules, and related tax reports. Being a Tax Accountant - Entry typically reports to a supervisor or a manager. May require a bachelor's degree. Being a Tax Accountant - Entry works on projects/matters of limited complexity in a support role. Work is closely managed. Working as a Tax Accountant - Entry typically requires 0-2 years of related experience.	57618	63900	70800	78900	86275
Tax Accountant - Entry	Los Angeles, CA	The Tax Accountant - Entry helps with the preparation of local, state and federal level returns to be submitted within specified tax deadlines. Assists in the maintenance and preparation of tax-related items to include tax records, tax returns, tax schedules, and related tax reports. Being a Tax Accountant - Entry typically reports to a supervisor or a manager. May require a bachelor's degree. Being a Tax Accountant - Entry works on projects/matters of limited complexity in a support role. Work is closely managed. Working as a Tax Accountant - Entry typically requires 0-2 years of related experience.	54182	60100	66600	74200	81119
Tax Accountant - Entry	Chicago, IL	The Tax Accountant - Entry helps with the preparation of local, state and federal level returns to be submitted within specified tax deadlines. Assists in the maintenance and preparation of tax-related items to include tax records, tax returns, tax schedules, and related tax reports. Being a Tax Accountant - Entry typically reports to a supervisor or a manager. May require a bachelor's degree. Being a Tax Accountant - Entry works on projects/matters of limited complexity in a support role. Work is closely managed. Working as a Tax Accountant - Entry typically requires 0-2 years of related experience.	50937	56400	62400	69600	76155
Tax Accountant - Entry	Houston, TX	The Tax Accountant - Entry helps with the preparation of local, state and federal level returns to be submitted within specified tax deadlines. Assists in the maintenance and preparation of tax-related items to include tax records, tax returns, tax schedules, and related tax reports. Being a Tax Accountant - Entry typically reports to a supervisor or a manager. May require a bachelor's degree. Being a Tax Accountant - Entry works on projects/matters of limited complexity in a support role. Work is closely managed. Working as a Tax Accountant - Entry typically requires 0-2 years of related experience.	49219	54500	60300	67200	73482
Tax Accountant - Entry	Phoenix, AZ	The Tax Accountant - Entry helps with the preparation of local, state and federal level returns to be submitted within specified tax deadlines. Assists in the maintenance and preparation of tax-related items to include tax records, tax returns, tax schedules, and related tax reports. Being a Tax Accountant - Entry typically reports to a supervisor or a manager. May require a bachelor's degree. Being a Tax Accountant - Entry works on projects/matters of limited complexity in a support role. Work is closely managed. Working as a Tax Accountant - Entry typically requires 0-2 years of related experience.	47401	52500	58100	64800	70900

Show 10 per page

Examples and Exercises

Go to Google Classroom assignment Lesson 17 “Beautiful Soup”

Tasks

- Follow instructions for the examples
- Work on the exercises (**due** by 10/1 6:00 PM)

Lesson 18
Candlestick Plot

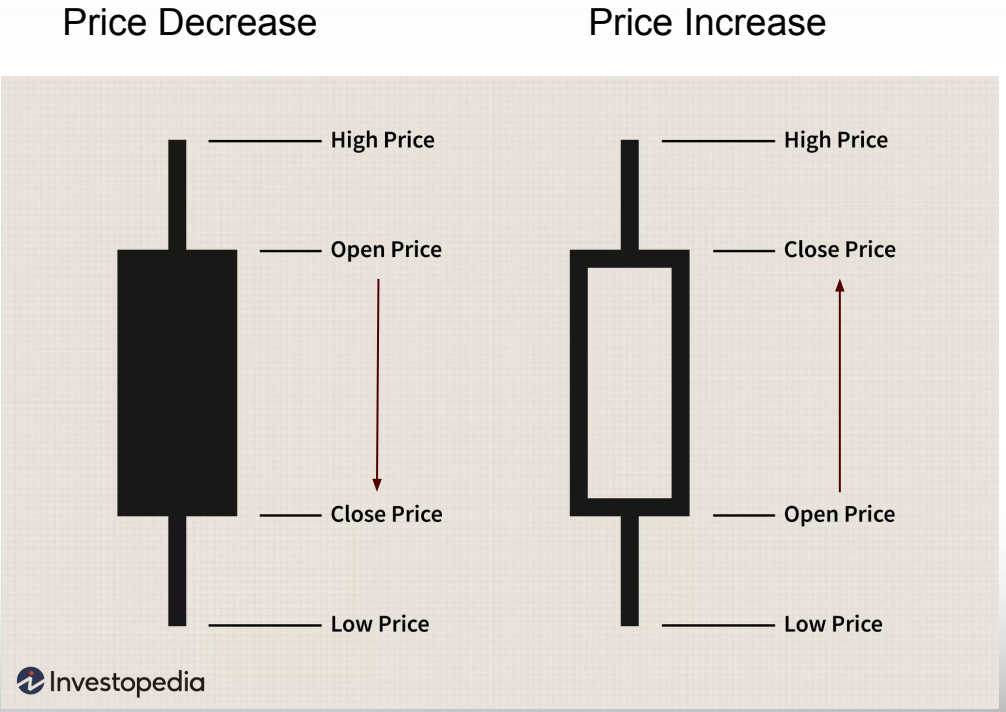
Use Matplotlib finance module (mplfinance) to plot stock price movement

Candlestick plot

Candlestick Plot

The color of candlestick rectangle area depends on the relationship between the opening and closing price of a stock.

The color of the rectangle shows increase/decrease



Candlestick Plot

Keyword arguments:

- type
- style
- volume (trade volume)
- mav (moving averages)

Examples and Exercises

Go to Google Classroom assignment Lesson 18 “Candlestick Plot”

Tasks

- Follow instructions for the examples
- Work on the exercises (**due** by 10/1 6:00 PM)

References

- https://www.tutorialspoint.com/beautiful_soup/beautiful_soup_quick_guide.htm
- http://www.compjour.org/warmups/govt-text-releases/intro-to-bs4-lxml-parsing-wh-press-briefings/#converting-html-text-into-a-data-object
- https://www.w3schools.com/html/
- https://github.com/matplotlib/mplfinance