CSC 4780/6780
Fall 2022
Homework 4

September 10, 2022

This homework is due at 11:59 pm on Sunday, Sept 18 It must be uploaded to iCollege by then. No credit will be given for late submissions. A solution will be released by noon on Monday, Sept 19.

Once again: it is a good idea to get this done and turn it in early.

1 Read

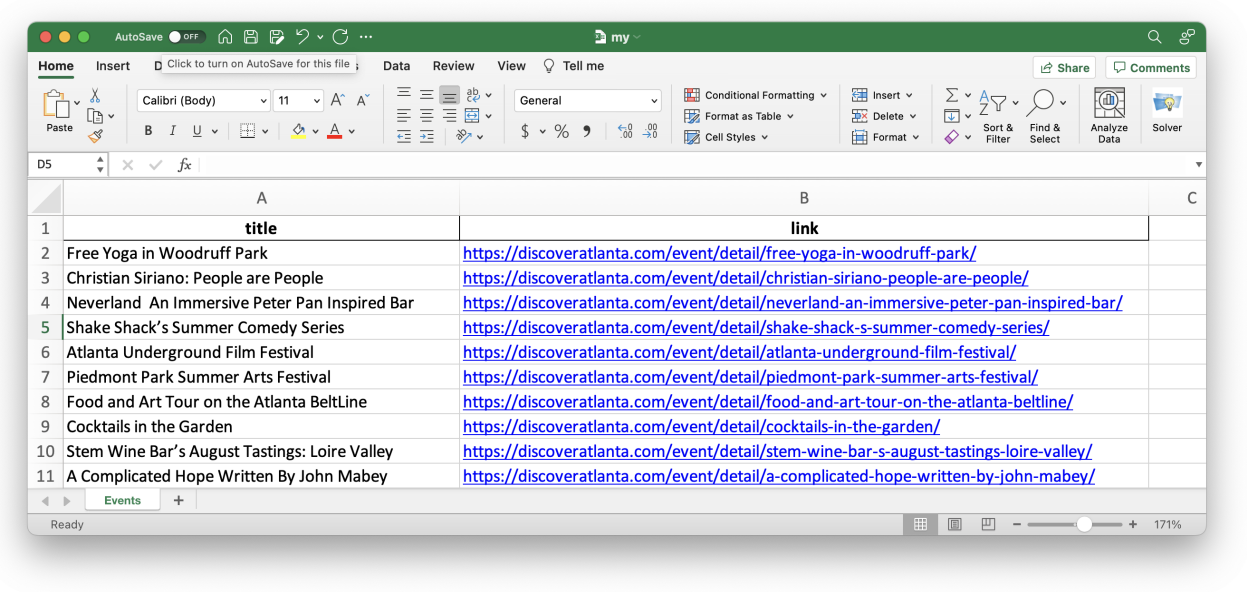
At this point you should have read up to the start of Chapter 8: *Probability, Distributions, and Sampling.*

2 Scrape a webpage

(6 points) Create a python program called `scrape.py` that takes a date in ISO format as an argument:

```
> python3  scrape.py 2022-10-02 result.xlsx
```

The program will then create and excel spreadsheet that lists the names of events that will happen on that date and their urls. It should look like this when you open it in Excel:



	A	B	C
1	title	link	
2	Free Yoga in Woodruff Park	https://discoveratlanta.com/event/detail/free-yoga-in-woodruff-park/	
3	Christian Siriano: People are People	https://discoveratlanta.com/event/detail/christian-siriano-people-are-people/	
4	Neverland An Immersive Peter Pan Inspired Bar	https://discoveratlanta.com/event/detail/neverland-an-immersive-peter-pan-inspired-bar/	
5	Shake Shack's Summer Comedy Series	https://discoveratlanta.com/event/detail/shake-shack-s-summer-comedy-series/	
6	Atlanta Underground Film Festival	https://discoveratlanta.com/event/detail/atlanta-underground-film-festival/	
7	Piedmont Park Summer Arts Festival	https://discoveratlanta.com/event/detail/piedmont-park-summer-arts-festival/	
8	Food and Art Tour on the Atlanta BeltLine	https://discoveratlanta.com/event/detail/food-and-art-tour-on-the-atlanta-beltline/	
9	Cocktails in the Garden	https://discoveratlanta.com/event/detail/cocktails-in-the-garden/	
10	Stem Wine Bar's August Tastings: Loire Valley	https://discoveratlanta.com/event/detail/stem-wine-bar-s-august-tastings-loire-valley/	
11	A Complicated Hope Written By John Mabey	https://discoveratlanta.com/event/detail/a-complicated-hope-written-by-john-mabey/	

Behind the scenes, your program will

- fetch the web page at `https://discoveratlanta.com/events/all/`
- parse the result using BeautifulSoup and `html.parser`
- step through each article inspecting the dates of the events
- skip articles that do not contain the desired date
- for articles that have the desired date, note the title and the URL
- make a dataframe with all the titles and URLs
- write the dataframe to an `ExcelWriter`
- resize the columns to be a reasonable width
- write it to the file named on the command line

You are putting data into only 2 columns – Don't include the dataframe's index in the excel file.

3 Analyze the residual from the last exercise

(4 points) My solution to last week's regression problem (`linreg_scikit.py` and `util.py`) are in this directory. Extended it to save a histogram of the residual as `res_hist.png`.

Extended `linreg_scikit.py` again to use scipy's `kstest` to confirm that the residual really resembles a normal distribution. The test returns a P-value; if the P-value is less than 0.05, you can assume the residual is normally distributed.

Now that you know it is a normal distribution, extend `linreg_scikit.py` yet again to print your confidence like this "68% of the estimates done with this formula will be within $89.12 of the correct price. 95% will be within $140.19 of the correct price."

4 What to turn in

If your name is Fred Jones, you will turn in a zip file called `HW04_Jones_Fred.zip` of a directory called `HW04_Jones_Fred`. It will contain:

- `scrape.py`
- `result.xlsx`
- `linreg_scikit.py`
- `util.py`
- `properties.xlsx`
- `res_hist.png`

Be sure to format your python code with black before you submit it.

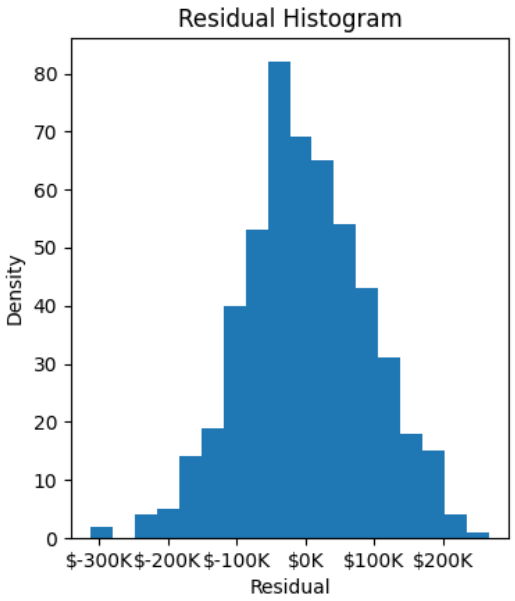
We will run your code like this:

```
cd HW04_Jones_Fred
python3 scrape.py 2022-10-02 result.xlsx
python3 linreg_scikit.py properties.xlsx
```

The output from the second program should look like this:

```
> python3 linreg_scikit.py properties.xlsx
Read 519 rows, 5 features from 'properties.xlsx'.
predicted price = $32,362.85 + ($85.61 x sqft_hvac) + ($2.73 x sqft_yard) +
    ($59,195.07 x bedrooms) + ($9,599.24 x bathrooms) +
    ($-17,421.84 x miles_to_school)
Kolmogorov-Smirnov: P-value = 4.154181404788638e-129
    The residual follows a normal distribution.
68% of predictions with this formula will be within $91,849.54 of the actual price.
95% of predictions with this formula will be within $183,699.08 of the actual price.
```

And should generate a histogram like this:



Do this work by yourself. Stackoverflow is OK. A hint from another student is OK. Looking at another student's code is *not* OK.

5 Extra help

Here is a nice tutorial on Beautiful Soup: `https://youtu.be/87Gx3U0BDlo`

Getting ahead: Soon we will be doing classification. Here is a good discussion of metrics for the quality of a classifier: `https://www.youtube.com/watch?v=8d3JbbSj-I8`