Take company names from wikipedia

You can directly access the htmls by going to:

Name = company name with periods and spaces replaced with “-”

Replace “, “ too

Replace Company with “ CO”

Replace corporation with “CORP”

Some companies may need add “Inc” to the end

Or plc

Or co

Or “the”

Or “new”

Or “corp”

replace & with “and”

## Apartment Investment & Management replace with APARTMENT INVST & MGMT CO

Management replace with “manage”

Some

Honestly, if we have an error, it’s probably easiest to just post the company name and come back later manually

<https://www1.salary.com/Name-Executive-Salaries.html>

I’ll get a table with a all the people

Find the row who contains the 4 “magic words”

CEO Name, CEO Pay, Median Employee Pay, and CEO Pay Ratio

Put it into a csv

Put it in the Pandas dataframe while prototyping

Adding ?year=2019 to the end allows for years, up to 2016

The latest I’ve seen the “magic words” appear is 2017 for home depot

I’m pretty sure 2019 and 2018 definitely has CEO Name, CEO Pay, Median Employee Pay, and CEO Pay Ratio

But while 2017 is possible, I wouldn’t count on it

Also some companies have some weird stuff lol

Marathon petroleum corp’s median income went up by like 100k from 2017 to 2018

If we scrape the other tables, look for “CEO” or “Chief Executive Officer”

Notes:

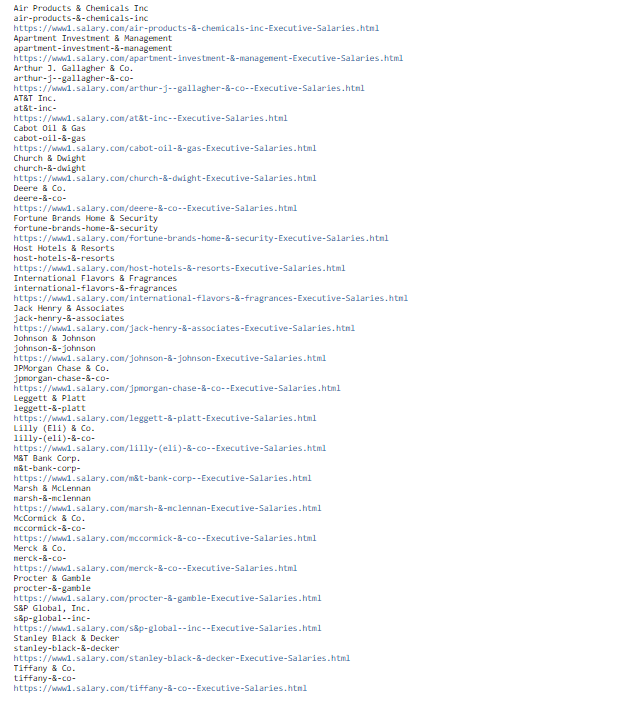
AT&T Inc.may be excluded?

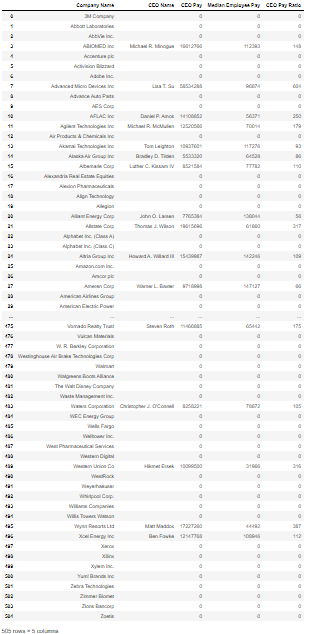
For some reason I can’t access the urls with years in them

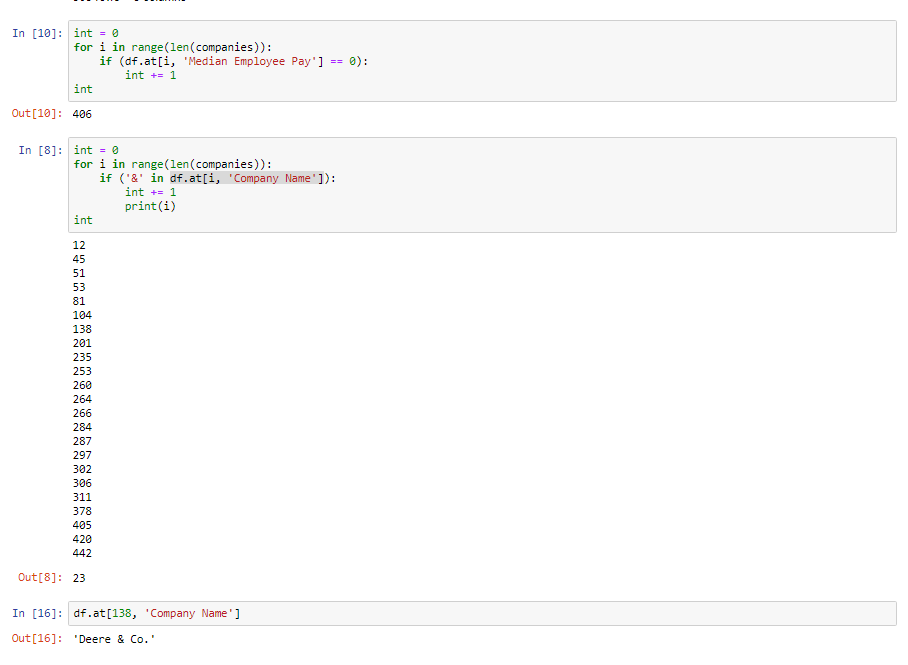
Fixed, use VPN

Chipotle compensation is wrong

Results for first go through of multi-site-ratio







2nd:

Disney has a really weird name (DISNEY-WALT-CO)

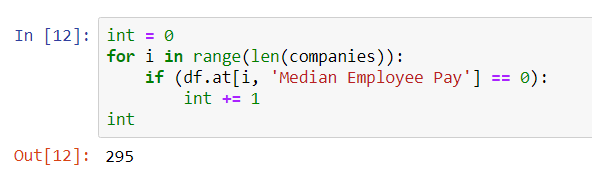
Accenture physically doesn’t have a ratio…

Neither does salesforce

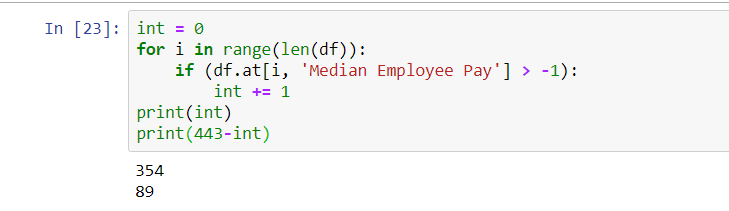
Neither does abott labs

AES corp ends with “the”

Did nearly triple the yield though



Try 3:



Iit got a connection error, but I think it’ll get everything that doesn’t have a terrible name