Intermediate Python for Data Science: CAPSTONE PROJECT

What makes a CEO stand out?

An exploratory analysis of American executives.

DENNIS GHELDOLF
May 1st, 2019

Project Motivation

A CEO walks into a bank..

.. and the bank wants to judge not only their business model, but also how this person stacks up in the industry.

Is this 28-year old really going to open a law office?
A college drop-out wants to start a technology company.
An award? Doesn't everyone have one of those?

Does any of this matter?

How Data Science helps

Challenge	Tool
Understanding from Executive Characteristics ('features').	EDA, inferential statistics and machine learning
Structured info on executives from biographies	Keyword rule-based logic, NLP
Biographies from names.	Web scraping
Getting names and titles from semi-structured forms.	Regex
Getting filings.	Download structured data from web
Storing all this data for processing.	Pandas

What data is available

Signatures at the end of every Quarterly and Annual SEC Filing:

Exhibit 31.1

CERTIFICATIONS

I, Robert J. Dennis, certify that:
[…]
Robert J. Dennis
Chief Executive Officer

Biographies on bloomberg.com for most public executives:

Lisa T. Su

President, CEO & Non-Independent Director,Advanced Micro Devices, Inc.

Age	Total Calculated Compensation	
48	**$13,356,392**	This person is connecte organizations across **4** c
	As of Fiscal Year 2018	See Board Relationships

Background*

Dr. Lisa T. Su has been the Chief Executive Officer and President August 10, 2014 and October 2014 respectively. Dr. Su served as Devices, Inc. from July 1, 2014 to October 2014. She served as S

Data Wrangling

WEB SCRAPING

```
In [50]: # Clean text of HTML codes and stray tags.
         def clean(text):
             return BeautifulSoup(text).get_text()

         def get_filings(URL):
             # Polite web scraping
             time.sleep(random.randint(1,3)/3)

             r = requests.get(URL)
             soup = BeautifulSoup(r.text)
```

- To obtain the Exhibits, in which execs sign with name and title.
- The URLs are a function of the filing's unique ID.
- BeautifulSoup cleans the page by stripping HTML tags and HTML character codes.

REGEX

REGULAR EXPRESSION — 1 match, 497 steps (~7ms)

```
r" I,.(.+?),?.certify.that:.+?(Chief.+?Officer|(?<!Vice )President(?! and Chief)) " gms
```

TEST STRING — SWITCH TO UNIT TESTS

```
I, Mimi E. Vaughn, certify that:
1. I have reviewed ...
... December 13, 2018
/s/ Mimi E. Vaughn
Mimi E. Vaughn
Chief Financial Officer
```

- Designed to extract name and title from Exhibits 31.1 / 31.2.
- Design and adapt to most common "exceptions" to the form:
  - ", Jr.", ", Sr.", ..
  - President vs. CEO.
  - Title in first line.

Data Wrangling

BIOGRAPHY INTERPRETATION

```
# Is it the correct Bio?
if not Name.lower() in Bio.lower():
    return [np.nan]*5

# Determine gender based on pronouns in text
gender = np.nan
if 'she' in Bio:
    gender = 'Female'
elif 'he' in Bio:
    gender = 'Male'
```

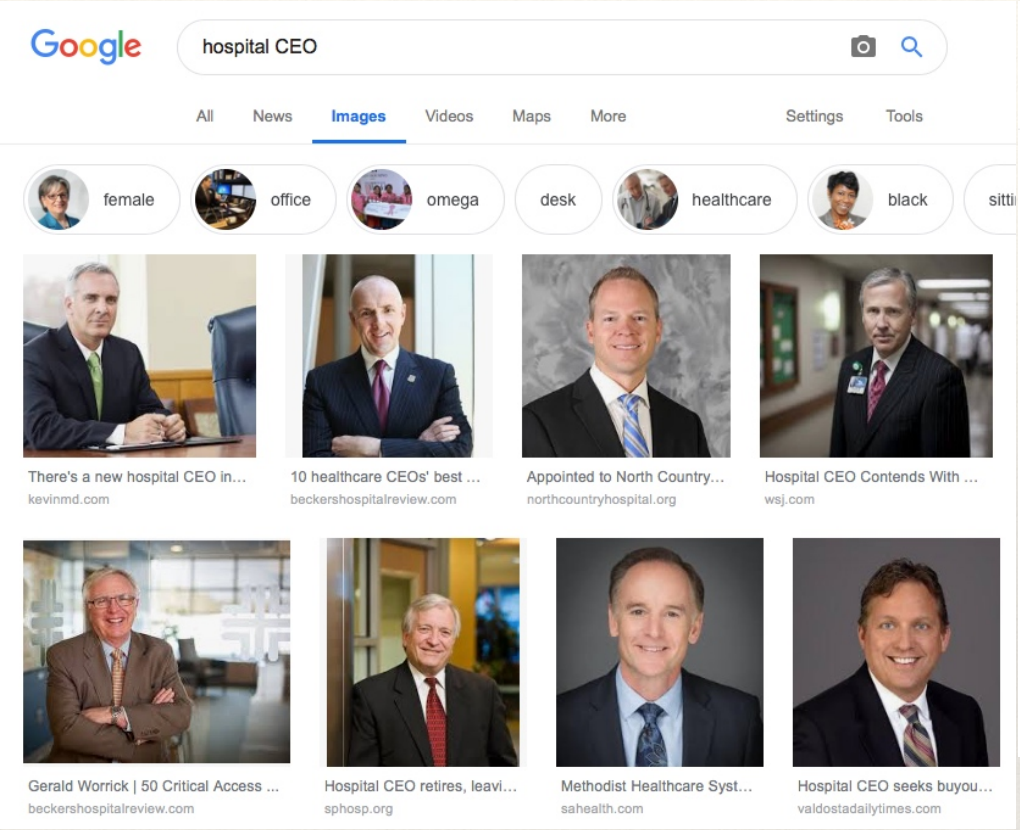
- Use rule-based logic to extract facts from text.
- Further optimizations:
  - Pre-process bios to strip punctuation and convert to lower case.
  - Incorporate NLTK.

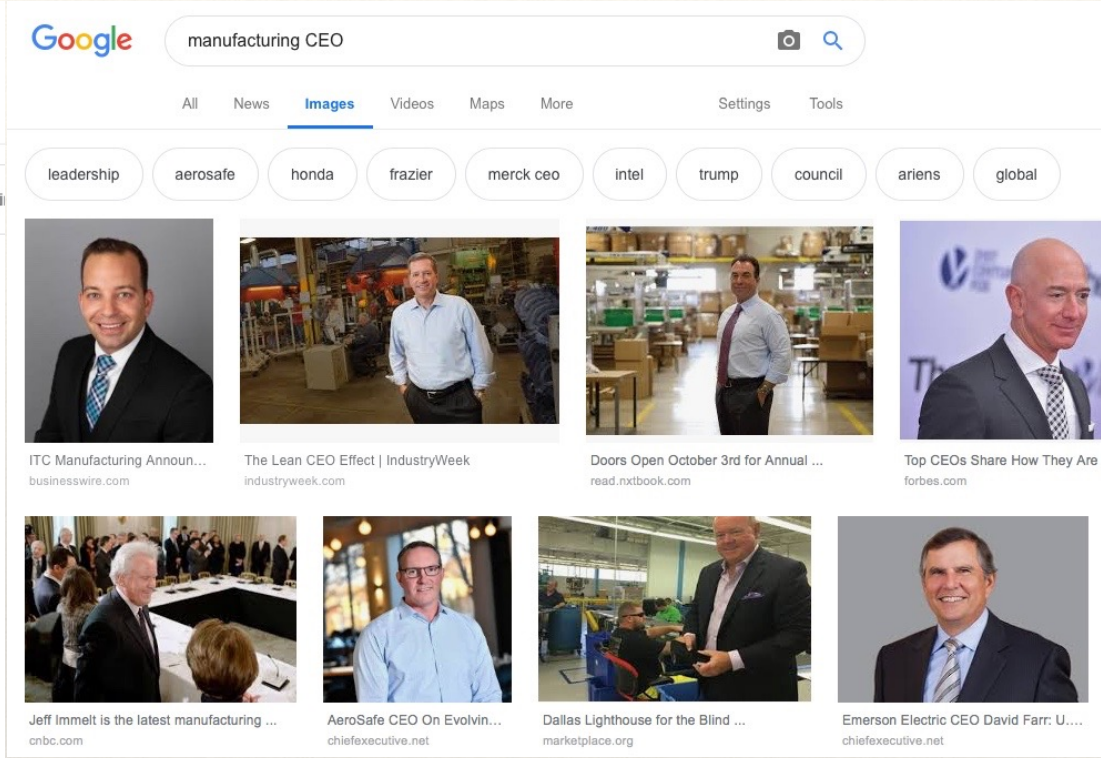
DATA HANDLING AND PASSING

```
for datafile in random.sample(os.li
    if not 'processed_'+datafile in
        df = pd.read_csv('./bios/'+
        df[['gender', 'degree', 'aw
        df.apply(features_from_bio,
        df['age'] = df.apply(verify
        df.to_csv('./bios_processed
```

- Many steps were time-intensive and subject to errors and interruptions.
- To deal with this, I processed data in chunks and wrote the output of to disk (CSV) as soon as it was done.
- HD5 may be cleaner?

EDA and Findings

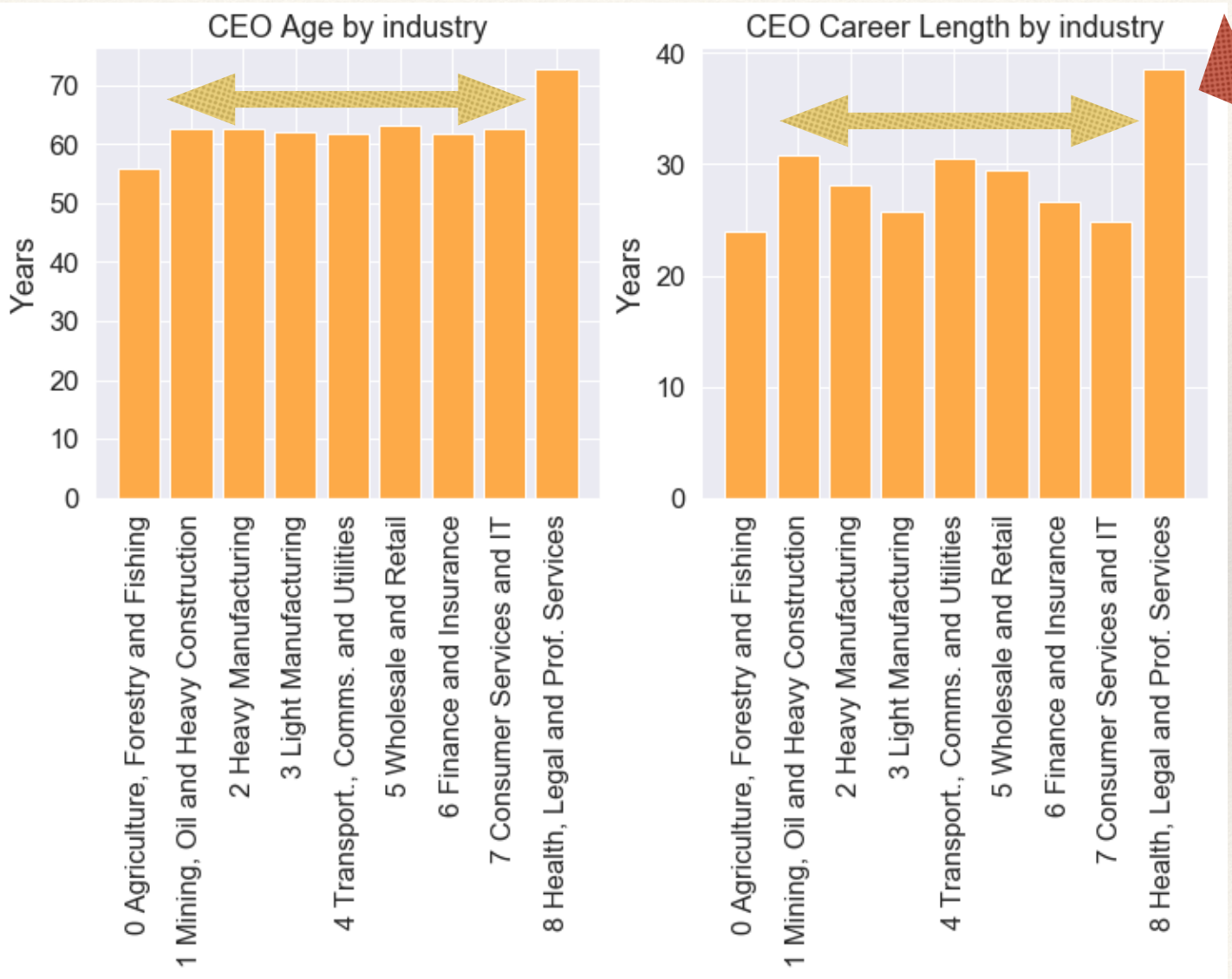




- Hospital: established in career.
- Manufacturing: groups into young and old.
- Two or three appear to be receiving an award.
- No women.

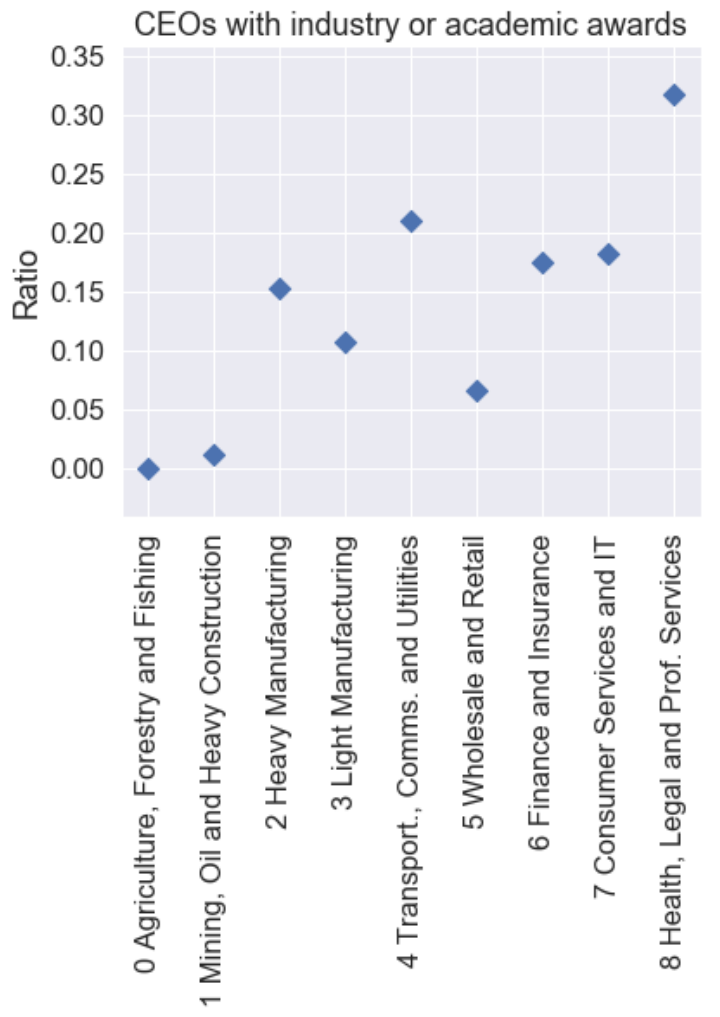
How many of these impressions can be generalized?

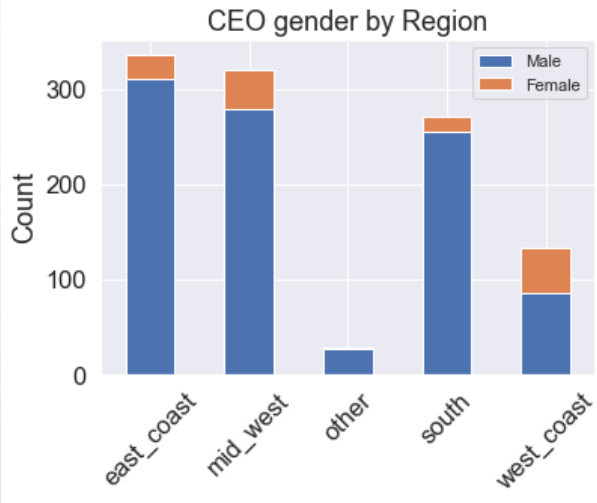
EDA and Findings

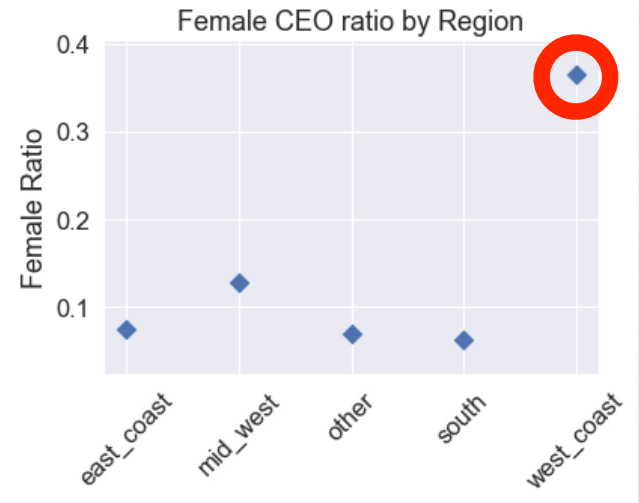


1. Age and Career Length in (8) is longer than in the other industries.
2. Despite similar ages, the career length in 1 to 7 varies. <u>Why is this?</u>

EDA and Findings

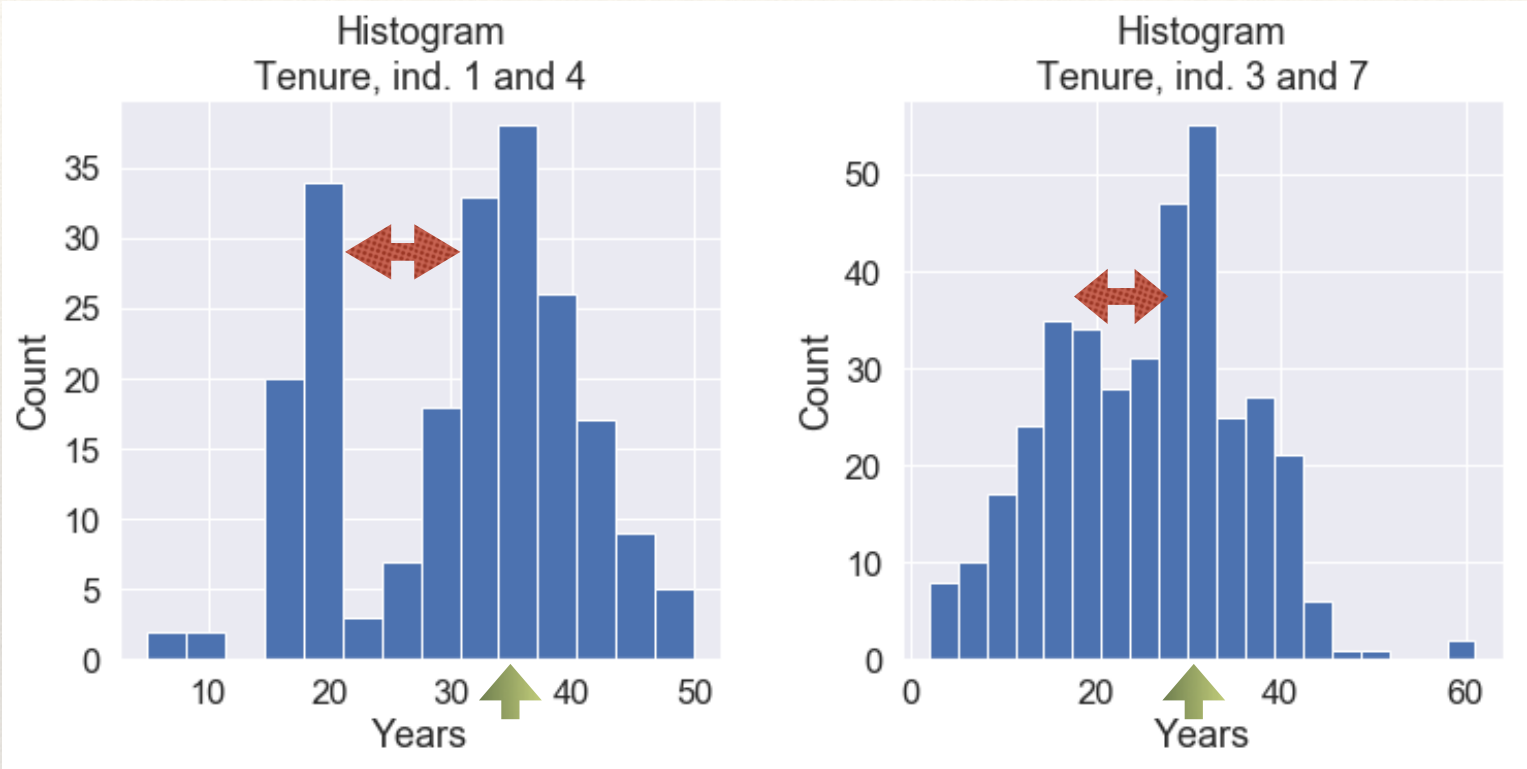






- 15 to 20% of CEOs hold an industry or academic award.
- About 10% of CEOs are women, except at the West Coast, where it is 35%.

In-depth analysis



- Career length is **BIMODAL**.
- Our original hunch are actually two problems:
  - Is there a significant difference between the means of the "later peaks" (ca. 35 and 30)? -> HYPOTHESIS TEST
  - What distinguishes the "Young CEOs" from the "Old CEOs"? -> CLASSIFICATION

In-depth analysis: hypothesis test

Hypothesis Test

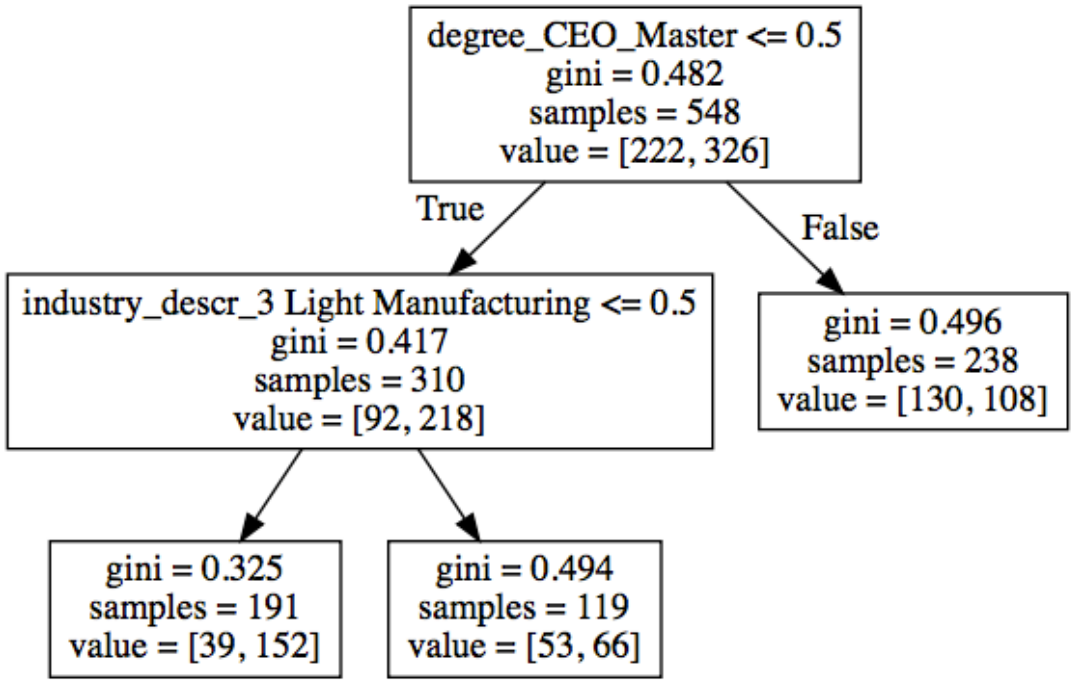
H0: There is no difference in mean Career Length of CEOs between industries (1, 4) and (3, 7) for Career Lengths over 25 yrs.

Method: an **independent two-sample t-test** is used to test the H0.

Result: the t-test returns
a p-value of 0.00 < 0.05.
We **Reject** the H0.
There exists a statistically significant difference between Mean Career Lengths Over 25 Yr between industries.

In-depth analysis: Decision Tree

	Features	Importance
6	degree_CEO_Master	0.662632
1	industry_descr_3 Light Manufacturing	0.337368



y = ‘long_tenured’

X = ['gender_CEO', 'industry_descr', 'top_univ_CEO', 'degree_CEO']

- The model was trained on the full data set, as the primary objective was to distinguish the most significant features.
- DecisionTreeClassifier() required “one-hot encoding” (with *pd.get_dummies(X,drop_first=True)*) for categoricals.

Results of Analysis

1. Average CEO Career Duration differs across industries, even when Average CEO Age does not.
2. A person is more likely to be a CEO before the 25th year of their career if they:
    a. Hold a Master's Degree, and
    b. Are active in Light Manufacturing

Interpretation

- *Manufacturing* and *Consumer Services & IT* attract CEOs that started their business career later in life.
- *Mining, Oil, Heavy Construction* and *Transportation, Communications and Utilities* support "Organization Man" lifetime career growth.

Recommendations

- CEOs in industries 0, 1, 2, 3 and 5 with an industry or academic award are rare (< 15%). Pay extra attention to them.
- Expect to see younger CEOs earlier in their careers, especially in *Consumer Services, IT and Light Manufacturing* and when they hold a *Master's Degree*.
- A CEO with a Master's Degree and active in Light Manufacturing that has been in his career for over 30 years is an anomaly. Understand their career path well.

Further improvements

- Refine the web scraping logic and text extraction logic to increase the sample size to several thousand.
- This would allow for higher granularity in the industry analysis.
- **Incorporate financial data to correlate executive features to financial performance.**
- Expand this analysis to CFOs.