Carol Chiu
April 1, 2020

Job Descriptions/ Skills Search Scraping Project

Objective:

The goal of this project is to search for the most common skillsets employers require for the average data science job in Los Angeles, CA and New York City, NY to determine which skills I need to learn to become more "marketable" for a data science role at a company.

Tools Used: BeautifulSoup, Python, pandas

Websites Used: Indeed.com, Glassdoor.com, Monster.com, LinkedIn.com

Data:

#Dataframe for scraped job postings where I counted the most common skillsets

Key_Skills	AWS	Alteryx	Athena	Azure	BI	Bayesian_modeling	Bayesian_statistics	C	C#	C++	...	stochastic	stochastic_modeling	supervised learning
Job_Posting1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0
Job_Posting2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0
Job_Posting4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting5	2.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting7	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting9	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting10	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0
Job_Posting11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting12	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	...	0.0	0.0	0.0
Job_Posting13	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0
Job_Posting15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting18	1.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	1.0	...	0.0	0.0	0.0
Job_Posting19	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting21	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0
Job_Posting22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting28	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting29	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Job_Posting30	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
Total_Count	6.0	3.0	1.0	2.0	1.0	1.0	3.0	4.0	1.0	3.0	...	1.0	1.0	1.0

31 rows × 121 columns

supervised_learning	time_series	trees	unstructured_data	unsupervised learning	unsupervised_learning	web2py
0.0	0.0	0.0	0.0	1.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	2.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	1.0
0.0	1.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	1.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
2.0	7.0	2.0	1.0	1.0	1.0	1.0

#Function where I searched for skills that occurred in over 20% of all the posts

```
for each in count:
    if each >= 6.0:
        print(each)
```

```
6.0
9.0
7.0
6.0
10.0
32.0
31.0
29.0
6.0
6.0
46.0
11.0
6.0
23.0
7.0
30.0
6.0
33.0
10.0
36.0
7.0
```

#The skills that show when searching for the most common words in data

```
df.columns[df.isin([6.0]).any()]
```

```
Index(['AWS', 'Hive', 'Spark', 'Tableau', 'Unnamed: 0',
       'artificial_intelligence', 'forecasting'],
      dtype='object')
```

```
df.columns[df.isin([7.0]).any()]
```

```
Index(['Excel', 'Unnamed: 0', 'data_mining', 'machine_learning', 'statistics',
       'time_series'],
      dtype='object')
```

```
df.columns[df.isin([9.0]).any()]
```

```
Index(['Cloud', 'Unnamed: 0'], dtype='object')
```

```
df.columns[df.isin([10.0]).any()]
```

```
Index(['Matlab', 'predictive_modeling'], dtype='object')
```

```
df.columns[df.isin([11.0]).any()]
```

```
Index(['algorithms'], dtype='object')
```

```
df.columns[df.isin([23.0]).any()]
```

```
Index(['data_analytics'], dtype='object')
```

```
df.columns[df.isin([29.0]).any()]
```

```
Index(['SQL'], dtype='object')
```

```
df.columns[df.isin([31.0]).any()]
```

```
Index(['R'], dtype='object')
```

```
df.columns[df.isin([32.0]).any()]
```

```
Index(['Python'], dtype='object')
```

```
df.columns[df.isin([33.0]).any()]
```

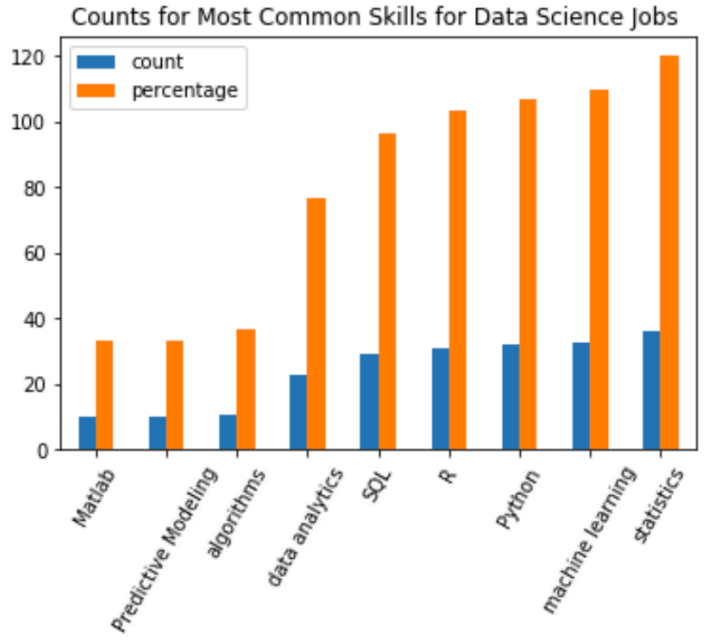
```
Index(['machine_learning'], dtype='object')
```

```
df.columns[df.isin([36.0]).any()]
```

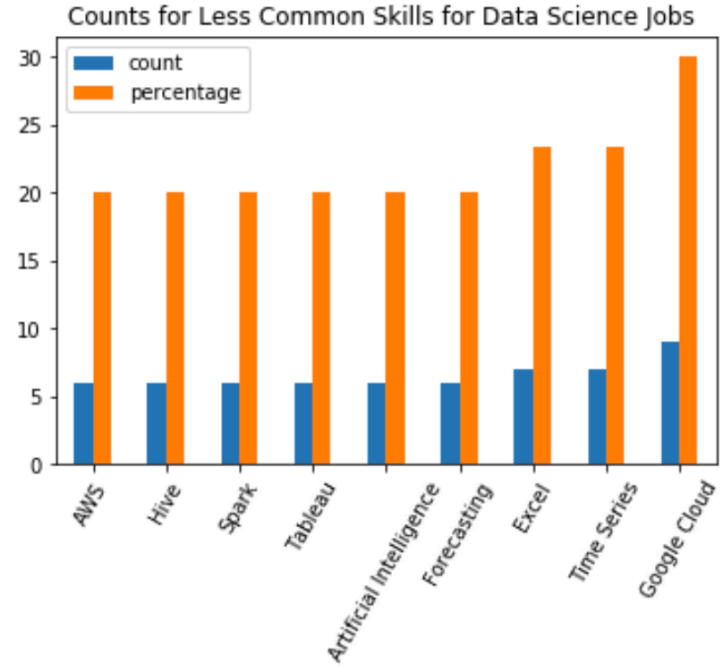
```
Index(['statistics'], dtype='object')
```

Data Plots/Results:

#The most common skillsets in the data science jobs (appeared in over 33% of searches on the website postings)



#The less common skillsets in the data science jobs (appeared in less than 33% of searches on the website postings)



Summary:

The skills that are most common in data science jobs (occur over 33% of postings) in Los Angeles, CA and New York City, NY are Matlab, predictive modeling, algorithms, data analytics, SQL, R, Python, machine learning, and statistics, most of which I am able to perform currently. The less common skills (under 33%) are AWS, Hive, Spark, Tableau, Artificial Intelligence, Forecasting, Excel, time series and Google Cloud. Of these remaining skillsets, I choose to learn AWS, Spark, time series, and Power BI in the next upcoming project.

[Credit & Special Thanks to: Alex Kwan, data analytics/data science mentor]