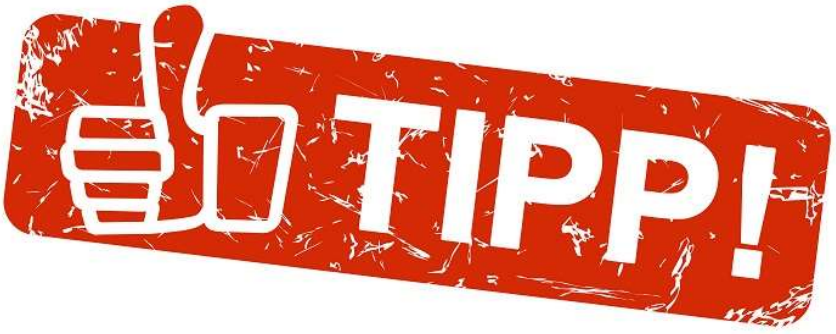
CASE STUDY 013

[Python]

US Working Visa Analysis



Here are some clues in case you are stuck with the case study:

1. The read_csv function could need to specify the encoding parameter to deal with these data sets:

```
encoding = 'latin1'
```

2. Consider both status 'CERTIFIED', 'CERTIFIED-WITHDRAWN' as certified visa.

3. To group, count and sort descending you can use this single statement:

```
df
[['SOC_CODE','SOC_NAME']].groupby('SOC_NAME').agg('count').sor
t_values(by='SOC_CODE', ascending=False)
```

You can make those steps one by one as well.

4. To filter data scientists, consider the following condition:

```
JOB_TITLE == 'DATA SCIENTIST'
```

5. You could make a merge with different column names, like:

```
pd.merge(visas, states, left_on = 'EMPLOYER_STATE', right_on =
'State abbreviation')
```

6. The top 5 percentile is equal to the 95th percentile. To calculate the 95th percentile for salary for instance, use the following code:

```
np.percentile(a = visas.PREVAILING_WAGE, q = 95)
```

7. To discover the industry you must make a join with the naics code data set. This data set has duplicated values and less than 1% of the visas data sets has an invalid naics code. You can deal with duplicated values before to load into the data frame.