dataclean

March 16, 2023

[5]:
```
import pandas as pd
import numpy as np
```

[9]:
```
# load data...source: nflcombineresults.com
path = '/Users/dylanjorling/UCLA/412proj/data/'
name = 'combine.csv'
combine = pd.read_csv(path+name)
combine.head()
```

[9]:

	Year	Name	College	POS	Height (in)	Weight (lbs)	Hand Size (in)	Arm Length (in)	Wonderlic	40 Yard	Bench Press	Vert Leap (in)	Broad Jump (in)	Shuttle	3Cone	60Yd Shuttle
0	1987	Mike Adams	Arizona State	CB	69.8	198	8.50	30.50	NaN	4.42	13.0	32.0	118.0	4.60	NaN	11.91
1	1987	John Adickes	Baylor	C	74.8	266	10.25	30.00	NaN	4.97	25.0	26.5	103.0	4.60	NaN	NaN
2	1987	Tommy Agee	Auburn	FB	71.8	217	9.00	30.75	NaN	NaN	15.0	NaN	NaN	NaN	NaN	NaN
3	1987	David Alexander	Tulsa (OK)	C	75.0	279	10.50	32.75	NaN	5.13	22.0	27.5	105.0	4.33	NaN	NaN
4	1987	Lyneal Alston	Southern Mississippi	WR	72.1	202	10.00	33.00	NaN	4.64	7.0	32.0	114.0	4.52	NaN	11.85

[11]:
```
# clean draft data
draft_init = pd.read_csv(path+'1987.csv')
draft_init2 = draft_init.iloc[:, :6]
draft_init2['College'] = draft_init['College/Univ']
draft_ttl = pd.DataFrame(columns=draft_init2.columns)
for i in range(1987, 2023):
    file=path+str(i) + '.csv'
```

```
    draft = pd.read_csv(file)
    draft_int = draft.iloc[:, :6]
    draft_int['College'] = draft['College/Univ']
    draft_int['Year'] = np.repeat(i, draft_int.shape[0])
    draft_ttl = pd.concat([draft_ttl, draft_int], axis=0)
draft_ttl.reset_index(inplace=True)
draft_ttl.drop(columns="index", inplace=True)
draft_ttl['Player'] = [player[:-4] if "HOF" in player else player for player in␣
 ↪draft_ttl["Player"]] # clean out "HOF"
draft_ttl['Year'] = draft_ttl['Year'].astype("int")
draft_ttl.shape
```

[11]: (9558, 8)

[12]:
```
# clean college
combine['College'].fillna(value="No College", inplace=True) # clean nas
clean_name = []
for name in combine['College']:
    if (name[-1:] == ')') & (name != 'Miami (OH)'):
        clean_name.append(name[:-5])
    else:
        clean_name.append(name)
combine['College'] = clean_name
clean_name = []
for name in combine['College']:
    if name[-5:] == 'State':
        clean_name.append(name[:-5] + 'St.')
    else:
        clean_name.append(name)
combine['College'] = clean_name
combine['College'].replace({'Southern California': 'USC',
                            'Louisiana St.': 'LSU',
                            'Boston College': 'Boston Col.',
                            'Miami': 'Miami (FL)',
                            'Texas Christian': 'TCU',
                            'Frenso St.': 'Fresno St.',
                            'Southern Methodist': 'SMU',
                            'Southern Mississippi': 'Southern Miss',
                            'Nevada Las Vegas':'UNLV',
                            'Missouri St.': 'Missouri State',
                            'Middle Tennessee St.': 'Middle Tenn. St.',
                            'Eastern Kentucky': 'East. Kentucky',
                            'Louisiana-Monroe': 'La-Monroe',
                            'Cal-State Fullerton': 'Cal State-Fullerton',
                            'Arkansas-Pine Bluff': 'Ark-Pine Bluff',
                            'Eastern Washington': 'East. Washington',
                            'Citadel': 'The Citadel',
```

```
                           'Cal Poly': 'Cal Poly-San Luis Obispo',
                           'Tennessee-Chattanooga': 'Chattanooga',
                           'Northwestern St.': 'Northwestern St. (LA)',
                           'Brigham Young': 'BYU',
                           'Louisiana-Lafayette': 'Louisiana',
                           'Stephen F. Austin': 'S.F. Austin',
                           'Eastern Michigan': 'East. Michigan',
                           'Northwest Missouri St.': 'NW Missouri St.',
                           'Southeast Missouri St.': 'SE Missouri St.',
                           'Alabama-Birmingham': 'Ala-Birmingham',
                           'Tennessee-Martin': 'UT Martin',
                           'Central St.': 'Central State (OH)',
                           'Central Missouri': 'Central Missouri St.',
                           'Wisconsin-Whitewater': 'Wisconsin-Whitewater',
                           'Wisconsin-Steven\'s Point': 'Wisconsin-Stevens␣
 ↪Point',
                           'UC-Davis': 'UC Davis',
                           'Mississippi Valley St.': 'Miss. Valley St.',
                           'North Carolina Charlotte': 'Charlotte',
                           'Eastern New Mexico': 'East. New Mexico',
                           'East. Illinois': 'Eastern Illinois',
                           'Sonoma': 'Sonoma St.',
                           'Missouri Western': 'Missouri Western St.',
                           'Kutztown': 'Kutztown (PA)',
                           'Albany St.': 'Albany State (GA)',
                           'Albany': 'Albany (NY)',
                           'Central Connecticut': 'Central Connecticut St.',
                           'East Central': 'East Central (OK)',
                           'Concordia - St Paul': 'Concordia-St.Paul (MN)',
                           'Charleston': 'Charleston (WV)',
                           'Augustana': 'Augustana (SD)',
                           'Angelo St.': 'Angelo State (TX)',
                           'Southwest Minnesota': 'SW Minnesota',
                           'Western Ontario (Ca': 'Western Ontario',
                           'Case Western': 'Case Western Reserve',
                           'Wayne St.': 'Wayne State (NE)'}, inplace=True)
```

[7]:
```
in_combine = []
for college in draft_ttl['College'].value_counts().index[:300]:
    if college in list(combine['College'].unique()):
        in_combine.append(college)
    else:
        print(college + " Not in set")
print(len(in_combine))
# couldnt find matches for these
```

```
Indiana (PA) Not in set
```

```
Oregon Tech Not in set
East. Illinois Not in set
Knoxville Not in set
NW Oklahoma St. Not in set
Wisconsin-LaCrosse Not in set
Boston Univ. Not in set
Robert Morris Not in set
California (PA) Not in set
Long Beach CC Not in set
290
```

[8]:
```
# now check how many drafted player have unique names
print(len(draft_ttl['Player']), len(draft_ttl['Player'].unique())) # 208 repeat␣
 ↪names. add position condition by might not match
unique_names = draft_ttl['Player'].unique()
dups = list(draft_ttl['Player'][draft_ttl['Player'].duplicated()])
drafted_list = []
for i, name in enumerate(combine['Name']):
    if (name not in dups) & (name in list(draft_ttl['Player'])):
        idx = draft_ttl['Player'][draft_ttl['Player'] == name].index[0]
        drafted_list.append(draft_ttl['Pick'][idx])
    elif name in list(draft_ttl['Player']):
        year = combine['Year'][i]
        college = combine['College'][i]
        idx = draft_ttl['Player'][(draft_ttl['Player'] == name) &
                                  (draft_ttl['Year'] == year) &
                                  (draft_ttl['College'] == college)].index
        if len(idx) > 0:
            drafted_list.append(draft_ttl['Pick'][idx[0]])
        else:
            drafted_list.append('Can\'t Find')
    else:
        drafted_list.append('Can\'t Find')
```

```
9558 9350
```

[9]:
```
combine['Pick'] = drafted_list
combine.head()
```

[9]:

	Year	Name	College	POS	Height (in)	Weight (lbs)	\
0	1987	Mike Adams	Arizona St.	CB	69.8	198	
1	1987	John Adickes	Baylor	C	74.8	266	
2	1987	Tommy Agee	Auburn	FB	71.8	217	
3	1987	David Alexander	Tulsa	C	75.0	279	
4	1987	Lyneal Alston	Southern Miss	WR	72.1	202	

	Hand Size (in)	Arm Length (in)	Wonderlic	40 Yard	Bench Press	\

```
0            8.50           30.50        NaN     4.42         13.0
1           10.25           30.00        NaN     4.97         25.0
2            9.00           30.75        NaN      NaN         15.0
3           10.50           32.75        NaN     5.13         22.0
4           10.00           33.00        NaN     4.64          7.0

   Vert Leap (in)  Broad Jump (in)  Shuttle  3Cone  60Yd Shuttle        Pick
0            32.0            118.0     4.60    NaN         11.91  Can't Find
1            26.5            103.0     4.60    NaN           NaN         154
2             NaN              NaN      NaN    NaN           NaN  Can't Find
3            27.5            105.0     4.33    NaN           NaN         121
4            32.0            114.0     4.52    NaN         11.85  Can't Find
```

[10]:
```
# count total players with a pick value and compare to 9558 total players in␣
↪drafted list
print(combine[combine['Pick'] != 'Can\'t Find'].shape[0])
# check for missing drafted players
missing_players = []
for name in draft_ttl['Player']:
    if name not in list(combine['Name']):
        missing_players.append(name)
len(missing_players)
```

```
7821
```

[10]:
```
1870
```

[11]:
```
# try to id cases where different names/spellings used
pot_matches = {}
for player in missing_players:
    idx = draft_ttl['Player'][draft_ttl['Player'] == player].index[0]
    col = draft_ttl['College'][idx]
    year = draft_ttl['Year'][idx]
    pot_matches[player] = list(combine[(combine['College'] == col) &␣
↪(combine['Year'] == year)]['Name'])
pot_matches = {k: v for k, v in pot_matches.items() if v}
# there are some lower/upper case matches and a bunch of nicknames....will␣
↪filter for last names too
```

[12]:
```
# find matches on lowercase last name match in addition to college and year
new_pot_matches = {}
for k, v in pot_matches.items():
    last_name = k.split()[-1].lower()
    match_list = []
    for i in v:
        last_name_match = i.split()[-1].lower()
        if last_name_match == last_name:
```

```
            match_list.append(i)
    new_pot_matches[k] = match_list
new_pot_matches = {k: v for k,v in new_pot_matches.items() if v}
```

[13]:
```
# write down mismatches: 'Emmitt Smith': ['Cedric Smith'], 'Al Johnson': ['Ben␣
↪Johnson']
del new_pot_matches['Emmitt Smith']
del new_pot_matches['Al Johnson']
new_pot_matches
for k, v in new_pot_matches.items():
    idx = draft_ttl[draft_ttl['Player'] == k].index[0]
    pick = draft_ttl['Pick'][idx]
    x = v[0]
    idxc = combine[combine['Name'] == x].index[0]
    combine['Pick'][idxc] = pick
```

```
/var/folders/7r/4ts1_nhj4lz2ccky2m4dl_6h0000gn/T/ipykernel_55078/649930420.py:10
: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  combine['Pick'][idxc] = pick
```

[14]:
```
# print out new missing
print(combine[combine['Pick'] != 'Can\'t Find'].shape[0])
# print missing pick data for 2000 + draft picks:
print(combine[(combine['Pick'] != 'Can\'t Find') & (combine['Year'] >= 2000)].
↪shape[0])
print(draft_ttl[draft_ttl['Year'] >= 2000].shape[0])
```

```
8166
5332
5871
```

[15]:
```
# at this point I am content with the amount of pick data we got...some picks do␣
↪not go to combine
# we probably missed a couple but would be hard to get many more I think can say␣
↪*most* of "Cant find" when undrafted
combine['Pick'].replace({'Can\'t Find': 'undrafted'}, inplace=True)
print(combine.isna().sum())
# looking at the nas, immediately we should drop the wonderlic and 60yd shuffle␣
↪columns
combine = combine.drop(columns=['Wonderlic', '60Yd Shuttle'])
print()
print()
print(combine.isna().sum())
```

```
Year                    0
Name                    0
College                 0
POS                     0
Height (in)             0
Weight (lbs)            0
Hand Size (in)       1309
Arm Length (in)      1625
Wonderlic           13113
40 Yard              1424
Bench Press          3636
Vert Leap (in)       1885
Broad Jump (in)      1998
Shuttle              2841
3Cone                5506
60Yd Shuttle        10302
Pick                    0
dtype: int64
```

```
Year                    0
Name                    0
College                 0
POS                     0
Height (in)             0
Weight (lbs)            0
Hand Size (in)       1309
Arm Length (in)      1625
40 Yard              1424
Bench Press          3636
Vert Leap (in)       1885
Broad Jump (in)      1998
Shuttle              2841
3Cone                5506
Pick                    0
dtype: int64
```

[16]:
```
# clean positions
combine['POS'].value_counts()
combine.replace({'NT': 'DT',
                 'CB': 'DB',
                 'DL': 'DT',
                 'OT': 'OL',
                 'OLB': 'LB',
                 'OG': 'OL',
                 'ILB': 'LB',
                 'FS': 'DB',
```

```
                'SS': 'DB',
                'C': 'OL',
                'FB': 'RB',
                'S': 'DB',
                'EDG': 'DE'},
                inplace=True)

# now drop special teams positions: P, K, LS
combine = combine[(combine['POS'] != 'P') & (combine['POS'] != 'K') &␣
↪(combine['POS'] != 'LS')]
combine['POS'].value_counts()
```

[16]:
```
DB    2423
OL    2286
WR    1787
LB    1637
RB    1509
DE    1051
DT    1008
TE     768
QB     741
Name: POS, dtype: int64
```

[ ]:
```
combine
```

[17]:
```
# lets coucheck nt complete cases
complete_cases = combine.dropna()
print(complete_cases.shape) #5855 complete cases
print()
print(complete_cases['Pick'].value_counts()) #2428 undrafted vs 3427 drafted
```

```
(5843, 15)

undrafted    2421
89             23
119            20
64             20
150            20
              ...
320             2
259             1
258             1
257             1
321             1
Name: Pick, Length: 262, dtype: int64
```

[ ]:
```
complete_cases.head()
```

[19]:
```
combine.columns
```

[19]:
```
Index(['Year', 'Name', 'College', 'POS', 'Height (in)', 'Weight (lbs)',
       'Hand Size (in)', 'Arm Length (in)', '40 Yard', 'Bench Press',
       'Vert Leap (in)', 'Broad Jump (in)', 'Shuttle', '3Cone', 'Pick'],
      dtype='object')
```

[25]:
```
# rename cols
combine.rename(columns={'Year':'year', 'Name':'name', 'College':'college',
                        'POS': 'pos', 'Height (in)': 'height', 'Weight (lbs)':␣
 ↪'weight',
                        'Hand Size (in)': 'hand_size', 'Arm Length (in)':␣
 ↪'arm_length',
                        '40 Yard':'forty', 'Bench Press': 'bench', 'Vert Leap␣
 ↪(in)': 'vert',
                        'Broad Jump (in)': 'broad'})
```

[25]:

	year	name	college	pos	Height (in)	Weight (lbs)	\
0	1987	Mike Adams	Arizona St.	DB	69.80	198	
1	1987	John Adickes	Baylor	OL	74.80	266	
2	1987	Tommy Agee	Auburn	RB	71.80	217	
3	1987	David Alexander	Tulsa	OL	75.00	279	
4	1987	Lyneal Alston	Southern Miss	WR	72.10	202	
...	...	...	...	..	...	...	
13539	2022	Mykael Wright	Oregon	DB	70.50	173	
13540	2022	Devonte Wyatt	Georgia	DT	74.88	304	
13541	2022	Jalen Wydermyer	Texas A&M	TE	75.88	255	
13542	2022	Nick Zakelj	Fordham	OL	78.13	316	
13543	2022	Bailey Zappe	Western Kentucky	QB	72.50	215	

	Hand Size (in)	Arm Length (in)	40 Yard	bench	Vert Leap (in)	\
0	8.50	30.50	4.42	13.0	32.0	
1	10.25	30.00	4.97	25.0	26.5	
2	9.00	30.75	NaN	15.0	NaN	
3	10.50	32.75	5.13	22.0	27.5	
4	10.00	33.00	4.64	7.0	32.0	
...	...	...	...	...	...	
13539	9.00	30.50	4.57	NaN	NaN	
13540	9.88	32.63	4.77	NaN	29.0	
13541	9.75	33.13	NaN	NaN	NaN	
13542	9.88	32.88	5.13	27.0	28.5	
13543	9.75	31.38	4.88	NaN	30.0	

	Broad Jump (in)	Shuttle	3Cone	Pick
0	118.0	4.60	NaN	undrafted
1	103.0	4.60	NaN	154
2	NaN	NaN	NaN	119

```
3                105.0     4.33    NaN         121
4                114.0     4.52    NaN   undrafted
...                ...      ...    ...         ...
13539              NaN       NaN    NaN   undrafted
13540            111.0       NaN    NaN          28
13541              NaN       NaN    NaN   undrafted
13542            110.0     4.71   7.75         187
13543            109.0     4.40   7.19         137

[13210 rows x 15 columns]
```

[32]:
```
combine.columns = ['year', 'name', 'college', 'pos', 'height', 'weight',␣
↪'hand_size',
                   'arm_length', 'forty', 'bench', 'vert', 'broad_jump',␣
↪'shuttle',
                   '3cone', 'pick']
combine.head()
```

[32]:

	year	name	college	pos	height	weight	hand_size	arm_length	forty	bench	vert	broad_jump	shuttle	3cone	pick
0	1987	Mike Adams	Arizona St.	DB	69.8	198	8.50	30.50	4.42	13.0	32.0	118.0	4.60	NaN	undrafted
1	1987	John Adickes	Baylor	OL	74.8	266	10.25	30.00	4.97	25.0	26.5	103.0	4.60	NaN	154
2	1987	Tommy Agee	Auburn	RB	71.8	217	9.00	30.75	NaN	15.0	NaN	NaN	NaN	NaN	119
3	1987	David Alexander	Tulsa	OL	75.0	279	10.50	32.75	5.13	22.0	27.5	105.0	4.33	NaN	121
4	1987	Lyneal Alston	Southern Miss	WR	72.1	202	10.00	33.00	4.64	7.0	32.0	114.0	4.52	NaN	undrafted

[33]:
```
# save full dataset
combine.to_csv('full_combine_data')
```