In [1]:
```
#import all modules
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [2]:
```
#read the date file into a dataframe called NBA from a csv file called all_seasons.csv
#This csv file is in the same folder with this notebook file
NBA = pd.read_csv('all_seasons.csv')
```

In [3]:
```
#show the top 5 rows of the NBA dataframe
NBA.head(5)
```

Out[3]:

	Unnamed: 0	player_name	team_abbreviation	age	player_height	player_weight	college	country	draft_year	draft_round	...	pts
0	0	Dennis Rodman	CHI	36.0	198.12	99.790240	Southeastern Oklahoma State	USA	1986	2	...	5.7
1	1	Dwayne Schintzius	LAC	28.0	215.90	117.933920	Florida	USA	1990	1	...	2.3
2	2	Earl Cureton	TOR	39.0	205.74	95.254320	Detroit Mercy	USA	1979	3	...	0.8
3	3	Ed O'Bannon	DAL	24.0	203.20	100.697424	UCLA	USA	1995	1	...	3.7
4	4	Ed Pinckney	MIA	34.0	205.74	108.862080	Villanova	USA	1985	1	...	2.4

5 rows × 22 columns

In [4]:
```
#Shows column info of each column
NBA.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11145 entries, 0 to 11144
Data columns (total 22 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Unnamed: 0         11145 non-null  int64  
 1   player_name        11145 non-null  object 
 2   team_abbreviation  11145 non-null  object 
 3   age                11145 non-null  float64
 4   player_height      11145 non-null  float64
 5   player_weight      11145 non-null  float64
 6   college            11145 non-null  object 
 7   country            11145 non-null  object 
 8   draft_year         11145 non-null  object 
 9   draft_round        11145 non-null  object 
 10  draft_number       11145 non-null  object 
 11  gp                 11145 non-null  int64  
 12  pts                11145 non-null  float64
 13  reb                11145 non-null  float64
 14  ast                11145 non-null  float64
 15  net_rating         11145 non-null  float64
 16  oreb_pct           11145 non-null  float64
 17  dreb_pct           11145 non-null  float64
 18  usg_pct            11145 non-null  float64
 19  ts_pct             11145 non-null  float64
 20  ast_pct            11145 non-null  float64
 21  season             11145 non-null  object 
dtypes: float64(12), int64(2), object(8)
memory usage: 1.9+ MB
```

In [5]:
```
# Remove the "Unnamed: 0" Column since it is duplicate with the index column
NBA = NBA.drop('Unnamed: 0', axis = 1)
```

In [6]:
```
#show the top 5 rows of the NBA dataframe again, no "Unnamed: 0" column again this time, looks better.
NBA.head(5)
```

Out[6]:

	player_name	team_abbreviation	age	player_height	player_weight	college	country	draft_year	draft_round	draft_number	...	p
0	Dennis Rodman	CHI	36.0	198.12	99.790240	Southeastern Oklahoma State	USA	1986	2	27	...	5
1	Dwayne Schintzius	LAC	28.0	215.90	117.933920	Florida	USA	1990	1	24	...	2
2	Earl Cureton	TOR	39.0	205.74	95.254320	Detroit Mercy	USA	1979	3	58	...	0
3	Ed O'Bannon	DAL	24.0	203.20	100.697424	UCLA	USA	1995	1	9	...	3
4	Ed Pinckney	MIA	34.0	205.74	108.862080	Villanova	USA	1985	1	10	...	2

5 rows × 21 columns

In [7]:
```
#How many records in total
NBA["player_name"].count()
```

Out[7]:
```
11145
```

In [8]:
```
#How many teams in total
NBA['team_abbreviation'].nunique()
```

Out[8]:
```
36
```

In [9]:
```
#What are team abbreviations of each team
NBA['team_abbreviation'].unique()
```

Out[9]:
```
array(['CHI', 'LAC', 'TOR', 'DAL', 'MIA', 'HOU', 'LAL', 'ATL', 'MIL',
       'DEN', 'SEA', 'POR', 'VAN', 'NJN', 'BOS', 'IND', 'SAC', 'MIN',
       'PHI', 'ORL', 'SAS', 'PHX', 'DET', 'CHH', 'CLE', 'GSW', 'UTA',
       'WAS', 'NYK', 'MEM', 'NOH', 'CHA', 'NOK', 'OKC', 'BKN', 'NOP'],
      dtype=object)
```

In [10]:
```
# Oldest age among all player records
NBA['age'].max()
```

Out[10]:
```
44.0
```

In [11]:
```
#Who is the oldest player among all player records
NBA[NBA['age'] == NBA['age'].max()]['player_name']
```

Out[11]:
```
4820    Kevin Willis
Name: player_name, dtype: object
```

In [12]:
```
#Who is the youngest player among all player records, return all if there are more than one.
NBA[NBA['age'] == NBA['age'].min()]['player_name']
```

Out[12]:
```
78      Jermaine O'Neal
342         Kobe Bryant
4286       Andrew Bynum
Name: player_name, dtype: object
```

In [13]:
```
#Each country has how many player records
NBA['country'].value_counts()
```

Out[13]:
```
USA                    9410
France                  153
Canada                  140
Spain                    79
Brazil                   78
                        ...
Guinea                    1
Trinidad and Tobago       1
Sudan (UK)                1
Sudan                     1
Ghana                     1
Name: country, Length: 76, dtype: int64
```

In [14]:
```
#Average points of all player records
NBA['pts'].mean()
```

Out[14]:
```
8.126487213997295
```

In [15]:
```
#Count of players per season
NBA.groupby(by = 'season')['player_name'].nunique()
```

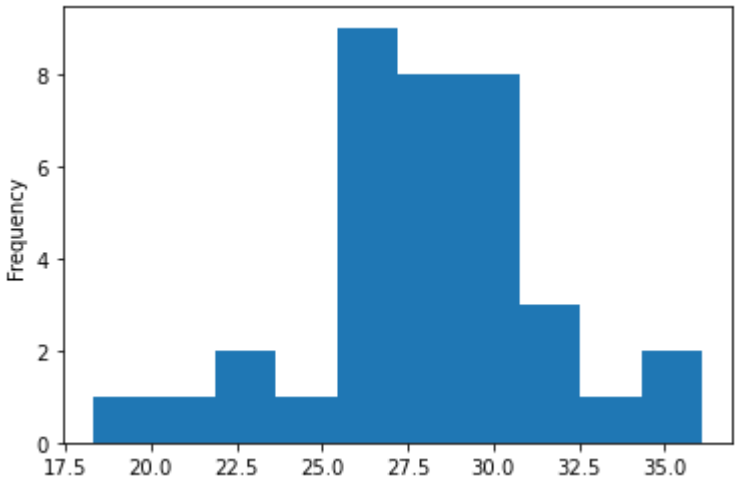
Out[15]:
```
season
1996-97    441
1997-98    439
1998-99    439
1999-00    438
2000-01    441
2001-02    440
2002-03    428
2003-04    442
2004-05    464
2005-06    458
2006-07    458
2007-08    450
2008-09    444
2009-10    442
2010-11    452
2011-12    478
2012-13    468
2013-14    481
2014-15    492
2015-16    476
2016-17    486
2017-18    540
2018-19    530
2019-20    514
Name: player_name, dtype: int64
```

In [16]:
```
#Each team's highest player average points
NBA.groupby(by = 'team_abbreviation')['pts'].max()
```

Out[16]:
```
team_abbreviation
ATL    29.4
BKN    27.4
BOS    28.9
CHA    25.6
CHH    26.8
CHI    29.6
CLE    31.4
DAL    28.4
DEN    28.9
DET    29.8
GSW    30.1
HOU    36.1
IND    25.8
LAC    26.9
LAL    35.4
MEM    21.1
MIA    30.2
MIL    29.6
MIN    26.5
NJN    25.2
NOH    22.9
NOK    18.3
NOP    28.1
NYK    28.7
OKC    32.0
ORL    32.1
PHI    33.0
PHX    26.6
POR    29.0
SAC    27.1
SAS    25.5
SEA    26.4
TOR    27.6
UTA    27.4
VAN    23.0
WAS    30.5
Name: pts, dtype: float64
```

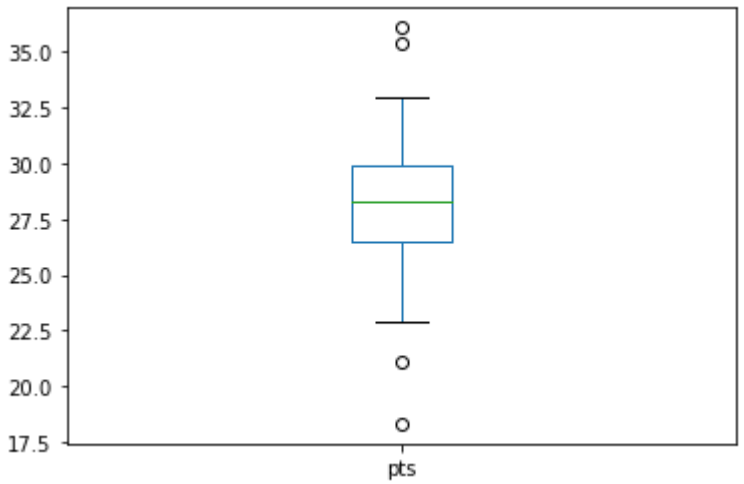
In [17]:
```
#A hist plot of Each team's highest player average points
NBA.groupby(by = 'team_abbreviation')['pts'].max().plot.hist()
```

Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x209566d1cd0>



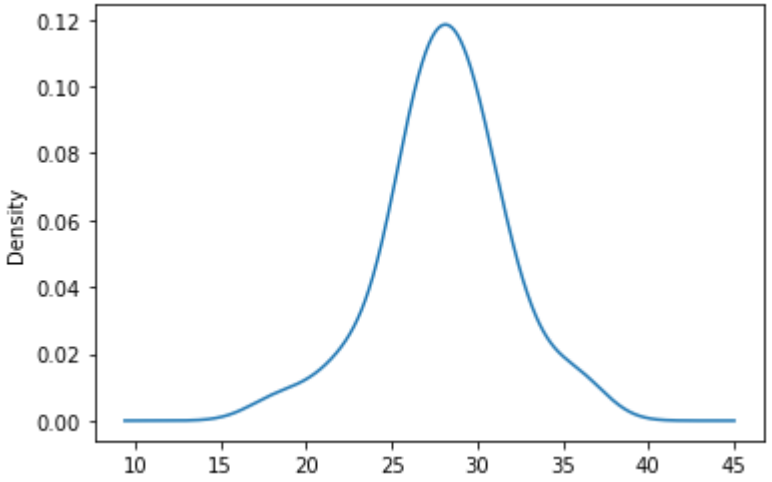
In [18]:
```
#A box plot of Each team's highest player average points
NBA.groupby(by = 'team_abbreviation')['pts'].max().plot.box()
```

Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x20956e2fd90>



In [19]:
```
#A kde plot of Each team's highest player average points
NBA.groupby(by = 'team_abbreviation')['pts'].max().plot.kde()
```

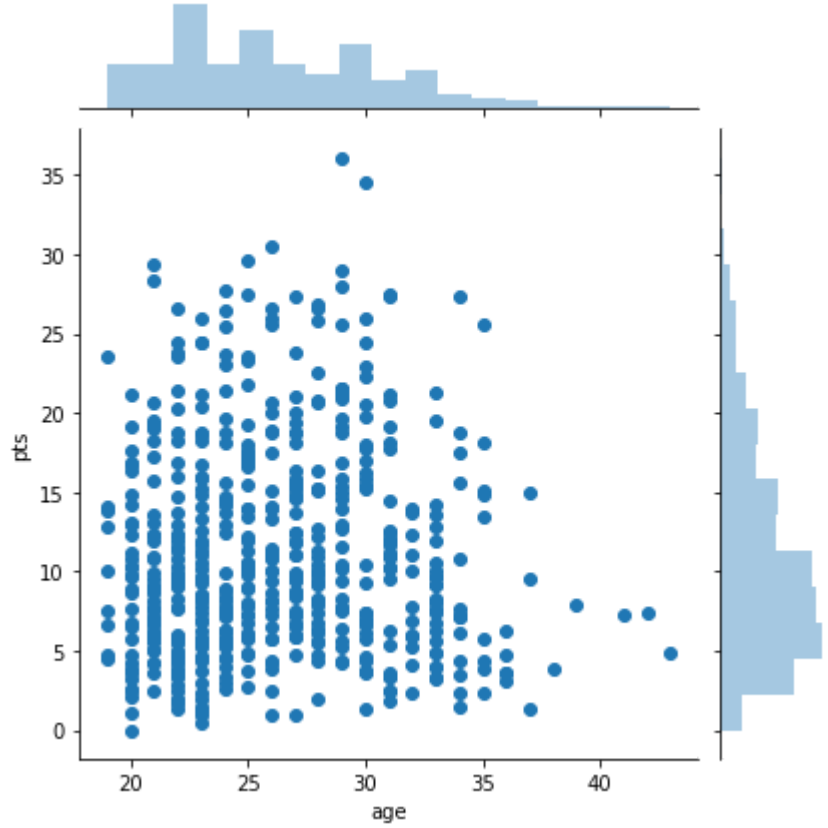
Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x20956ec1a90>



In [20]:
```
# Because the NBA dataframe has too many records, now I am generating another dataframe from the NBA dataframe.
#Get a sub dataframe of the NBA datafream where only season 2018-19 and 2019-20's 1st round drafted players' rec
NBA2 = NBA[((NBA['season'] == '2019-20') | (NBA['season'] == '2018-19')) & (NBA['draft_round'] == '1')]
```

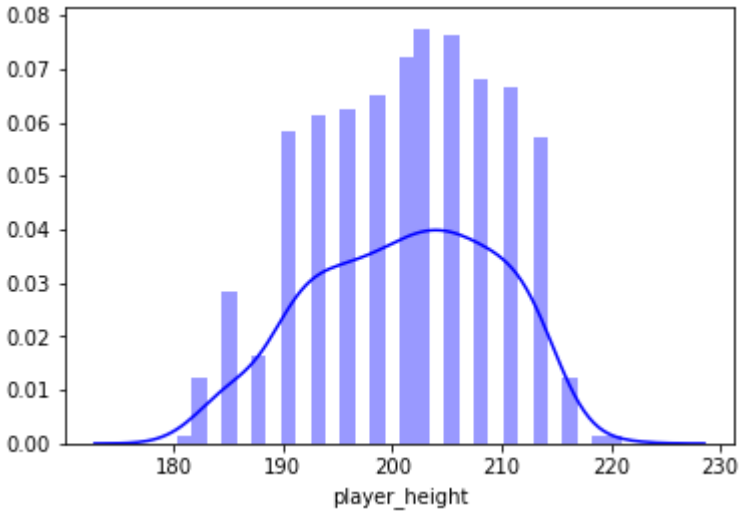
In [21]:
```
#Generate a jointplot betweeb age and pts of the NBA2 dataframe.
sns.jointplot(x='age', y = 'pts', data=NBA2)
```

Out[21]:
```
<seaborn.axisgrid.JointGrid at 0x20956f29d00>
```



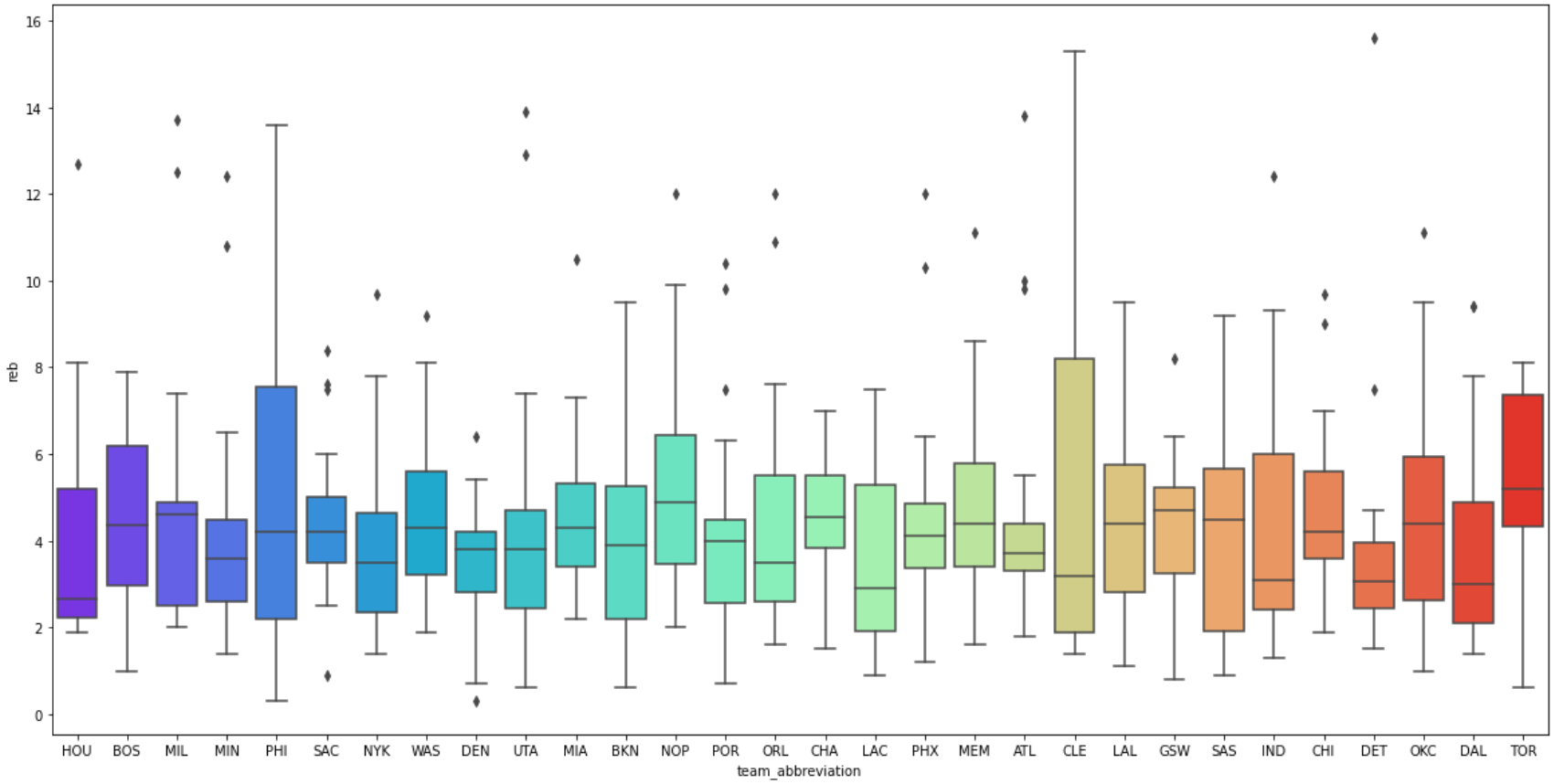
In [22]:
```
#A distplot of player height of the NBA2 dataframe, with the kde line.
sns.distplot(NBA2['player_height'], bins=30, kde=True, color='blue')
```

Out[22]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x20957081340>
```



In [23]:
```
#A boxplot showing relationship between each team and player's rebound stats, data frame is NBA2
plt.figure(figsize=(20,10))
sns.boxplot(x = 'team_abbreviation', y ='reb', data=NBA2, palette = 'rainbow')
```

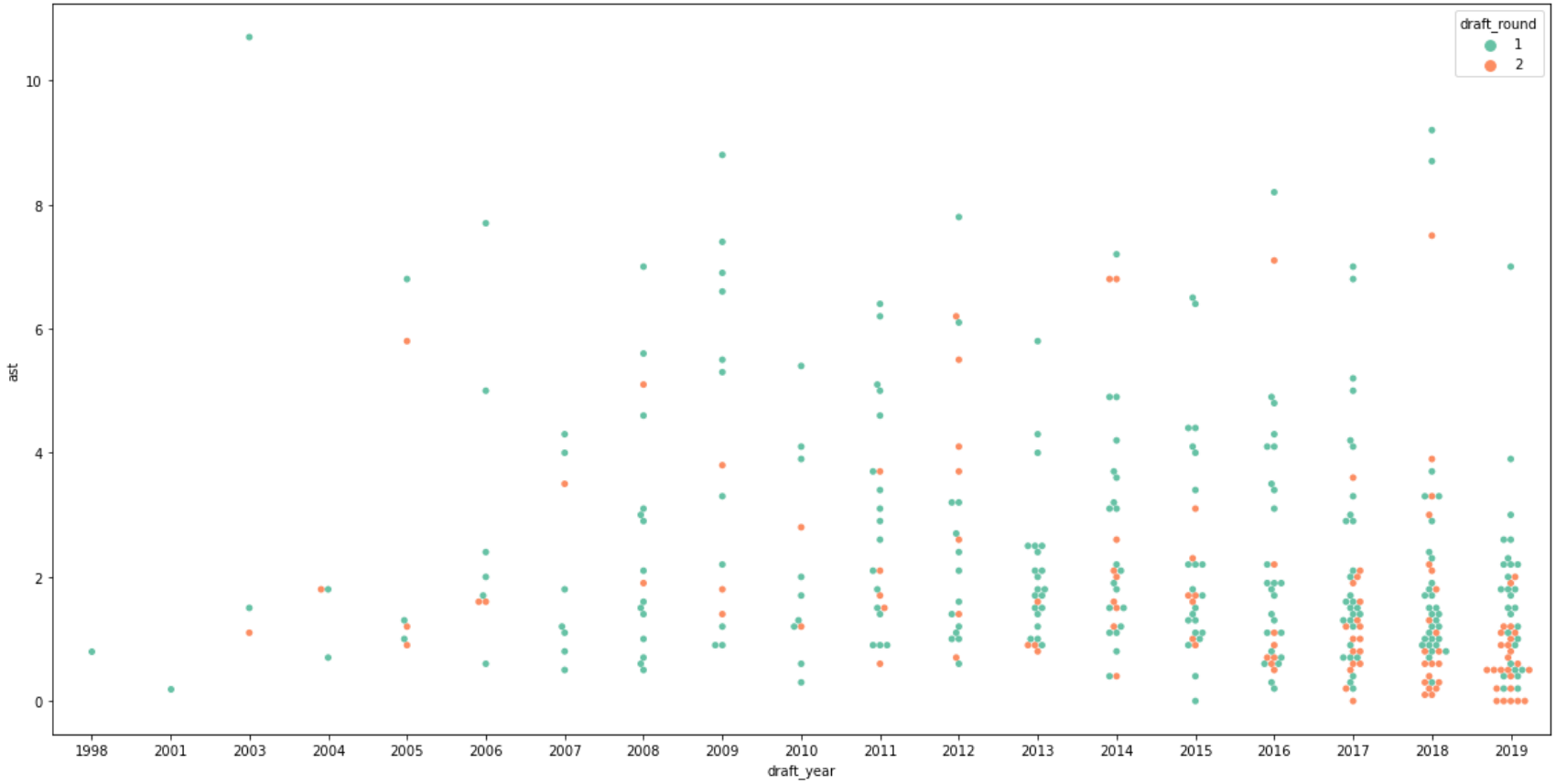
Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x2095711cc70>



In [24]:
```
# A sub dataframe NBA3 which has player records of season 2019-20 but only with players who were drafted in roun
NBA3 = NBA[(NBA['season'] == '2019-20')  & ((NBA['draft_round'] == '1') | (NBA['draft_round'] == '2'))]
```

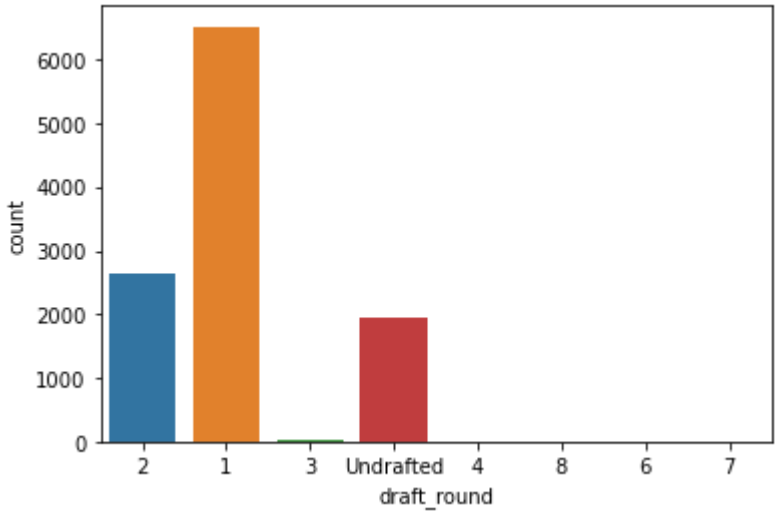
In [25]:
```
#A swarmplot showing relationship between draft years and player's assist stats, hue indicator is draft round,da
plt.figure(figsize=(20,10))
sns.swarmplot(x='draft_year',y='ast',data=NBA3,palette='Set2', hue='draft_round')
```

Out[25]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2095735d820>
```



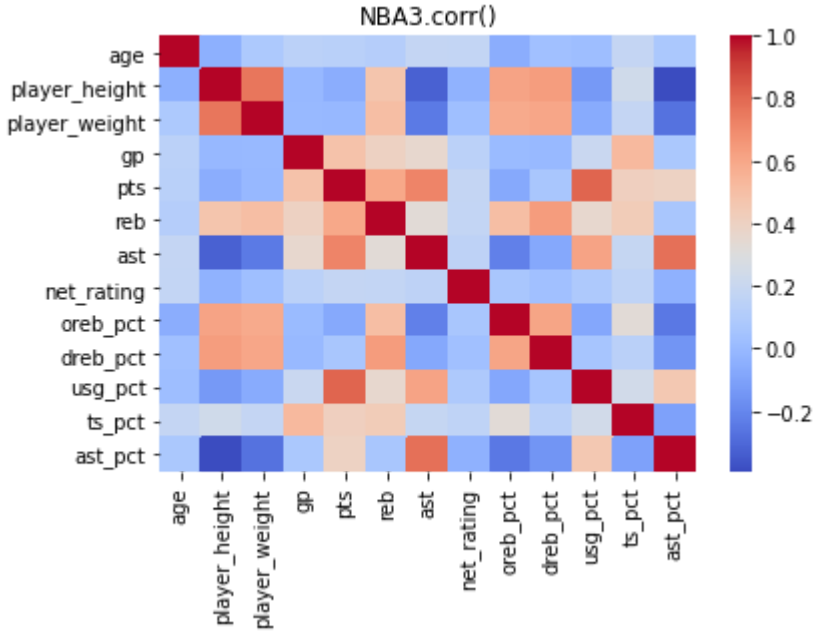
In [26]:
```
#count of player records of all times per round, dataframe is NBA
sns.countplot(x = 'draft_round', data=NBA)
```

Out[26]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x209570464f0>
```



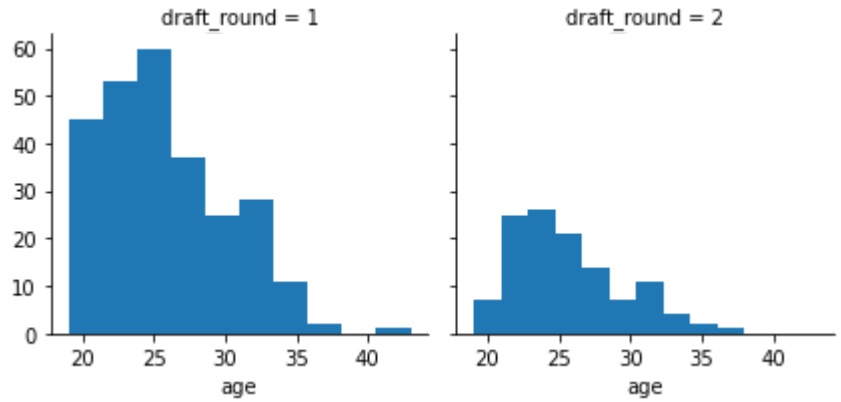
In [27]:
```
# Heat map of dataframe NBA3
sns.heatmap(NBA3.corr(), cmap='coolwarm')
plt.title('NBA3.corr()')
```

Out[27]:
```
Text(0.5, 1.0, 'NBA3.corr()')
```



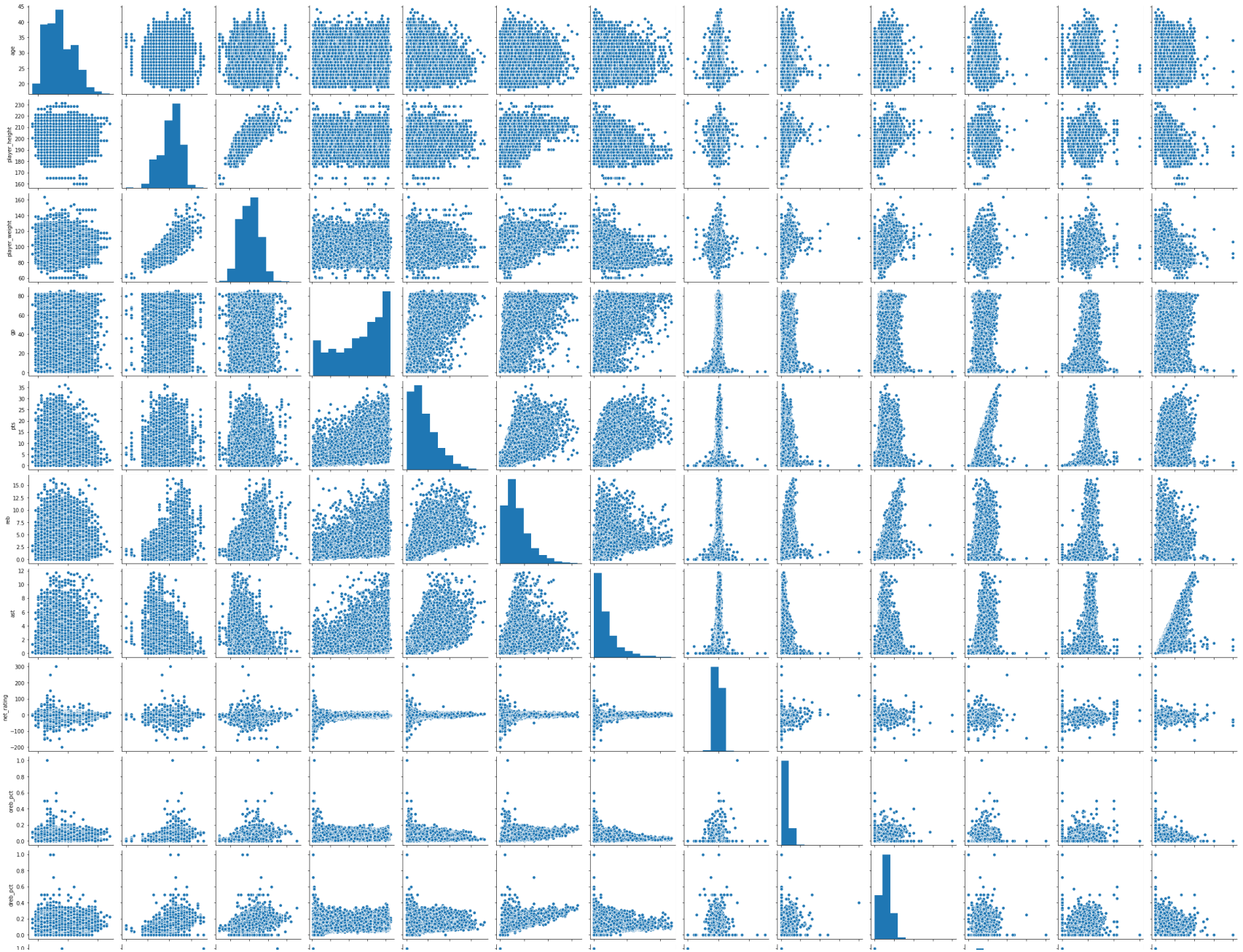
In [28]:
```
N = sns.FacetGrid(data=NBA3, col = 'draft_round')
N.map(plt.hist, 'age')
```

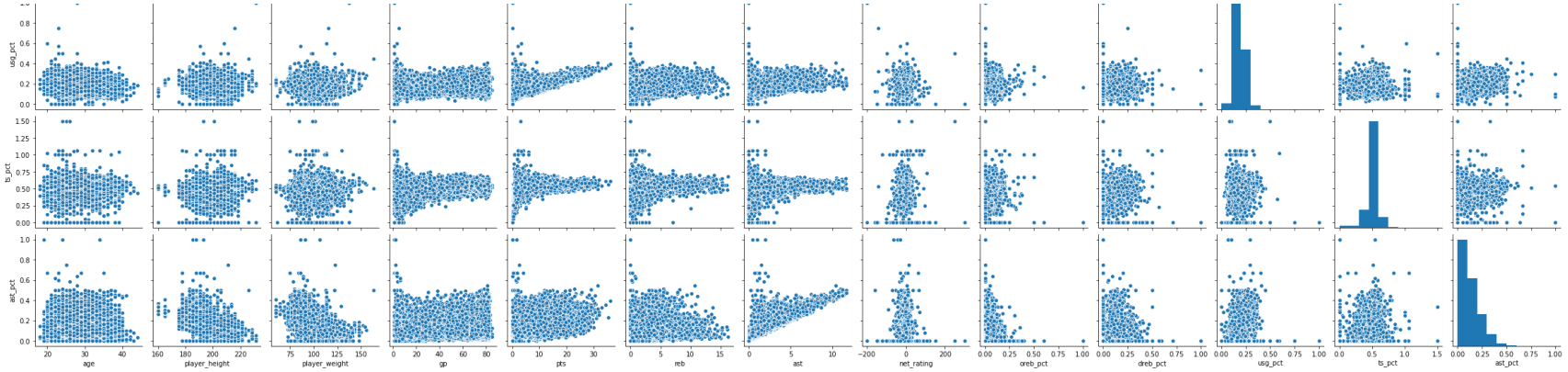
Out[28]: <seaborn.axisgrid.FacetGrid at 0x20957375820>



In [29]:
```
#Warning:  This may take a while to run since the dataframe NBA is big to generate a pairplot
sns.pairplot(data=NBA)
```

Out[29]: <seaborn.axisgrid.PairGrid at 0x20957c1e940>




usg_pct
ts_pct
ast_pct
age
player_height
player_weight
gp
pts
reb
ast
net_rating
oreb_pct
dreb_pct
usg_pct
ts_pct
ast_pct

In [ ]: