Salary Prediction using AMCAT dataset

1.0 Importing packages

In [72]:

```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn import preprocessing
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import GridSearchCV
from scipy.sparse import hstack
from sklearn.decomposition import PCA
from mpl_toolkits.mplot3d import Axes3D
from sklearn.metrics import accuracy_score
import warnings
warnings.filterwarnings("ignore")
```

1.1 Importing data

In [2]:

```
train = pd.read_excel('Data/train.xlsx')
test = pd.read_excel('Data/test.xlsx')
```

In [3]:

```
train.head()
```

Out[3]:

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10pe
0	train	203097	420000	2012-06-01	present	senior quality engineer	Bangalore	f	1990-02-19	
1	train	579905	500000	2013-09-01	present	assistant manager	Indore	m	1989-10-04	
2	train	810601	325000	2014-06-01	present	systems engineer	Chennai	f	1992-08-03	
3	train	267447	1100000	2011-07-01	present	senior software engineer	Gurgaon	m	1989-12-05	
4	train	343523	200000	2014-03-01	2015-03-01 00:00:00	get	Manesar	m	1991-02-27	

5 rows × 39 columns

In [4]:

```
test.head()
```

Out[4]:

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percenta
0	test	664736	?	?	?	?	?	m	1992-01-16	75
1	test	1123290	?	?	?	?	?	m	1992-06-05	83
2	test	1062444	?	?	?	?	?	f	1992-11-22	85
3	test	1072028	?	?	?	?	?	f	1990-10-17	81
4	test	267259	?	?	?	?	?	m	1990-03-20	78

5 rows × 39 columns

1.2 Understanding high level details about data

1.2.1 Missing values

In [5]:

```
train.isnull().sum()
```

Out[5]:

```
Unnamed: 0                0
ID                        0
Salary                    0
DOJ                       0
DOL                       0
Designation               0
JobCity                   0
Gender                    0
DOB                       0
10percentage              0
10board                   0
12graduation              0
12percentage              0
12board                   0
CollegeID                 0
CollegeTier               0
Degree                    0
Specialization            0
collegeGPA                0
CollegeCityID             0
CollegeCityTier           0
CollegeState              0
GraduationYear            0
English                   0
Logical                   0
Quant                     0
Domain                    0
ComputerProgramming       0
ElectronicsAndSemicon     0
ComputerScience           0
MechanicalEngg            0
ElectricalEngg            0
TelecomEngg               0
CivilEngg                 0
conscientiousness         0
agreeableness             0
extraversion              0
nueroticism               0
openess_to_experience     0
dtype: int64
```

In [6]:

```
test.isnull().sum()
```

Out[6]:

```
Unnamed: 0                0
ID                        0
Salary                    0
DOJ                       0
DOL                       0
Designation               0
JobCity                   0
Gender                    0
DOB                       0
10percentage              0
10board                   0
12graduation              0
12percentage              0
12board                   0
CollegeID                 0
CollegeTier               0
Degree                    0
Specialization            0
collegeGPA                0
CollegeCityID             0
CollegeCityTier           0
CollegeState              0
GraduationYear            0
English                   0
Logical                   0
Quant                     0
Domain                    0
ComputerProgramming       0
ElectronicsAndSemicon     0
ComputerScience           0
MechanicalEngg            0
ElectricalEngg            0
TelecomEngg               0
CivilEngg                 0
conscientiousness         0
agreeableness             0
extraversion              0
nueroticism               0
openess_to_experience     0
dtype: int64
```

There are no missing values in the training data and testing data.

1.2.2 Other information

In [7]:

```
train.describe().T
```

Out[7]:

	count	mean	std	min	25%	
ID	3998.0	663794.540520	363218.245829	11244.0000	334284.250000	6
Salary	3998.0	307699.849925	212737.499957	35000.0000	180000.000000	3
10percentage	3998.0	77.925443	9.850162	43.0000	71.680000	
12graduation	3998.0	2008.087544	1.653599	1995.0000	2007.000000	
12percentage	3998.0	74.466366	10.999933	40.0000	66.000000	
CollegeID	3998.0	5156.851426	4802.261482	2.0000	494.000000	
CollegeTier	3998.0	1.925713	0.262270	1.0000	2.000000	
collegeGPA	3998.0	71.486171	8.167338	6.4500	66.407500	
CollegeCityID	3998.0	5156.851426	4802.261482	2.0000	494.000000	
CollegeCityTier	3998.0	0.300400	0.458489	0.0000	0.000000	
GraduationYear	3998.0	2012.105803	31.857271	0.0000	2012.000000	
English	3998.0	501.649075	104.940021	180.0000	425.000000	
Logical	3998.0	501.598799	86.783297	195.0000	445.000000	
Quant	3998.0	513.378189	122.302332	120.0000	430.000000	
Domain	3998.0	0.510490	0.468671	-1.0000	0.342315	
ComputerProgramming	3998.0	353.102801	205.355519	-1.0000	295.000000	
ElectronicsAndSemicon	3998.0	95.328414	158.241218	-1.0000	-1.000000	
ComputerScience	3998.0	90.742371	175.273083	-1.0000	-1.000000	
MechanicalEngg	3998.0	22.974737	98.123311	-1.0000	-1.000000	
ElectricalEngg	3998.0	16.478739	87.585634	-1.0000	-1.000000	
TelecomEngg	3998.0	31.851176	104.852845	-1.0000	-1.000000	
CivilEngg	3998.0	2.683842	36.658505	-1.0000	-1.000000	
conscientiousness	3998.0	-0.037831	1.028666	-4.1267	-0.713525	
agreeableness	3998.0	0.146496	0.941782	-5.7816	-0.287100	
extraversion	3998.0	0.002763	0.951471	-4.6009	-0.604800	
nueroticism	3998.0	-0.169033	1.007580	-2.6430	-0.868200	
openess_to_experience	3998.0	-0.138110	1.008075	-7.3757	-0.669200	

In [8]:

```
test.describe().T
```

Out[8]:

	count	mean	std	min	25%	
ID	1500.0	665286.300667	360532.421901	7474.0000	335062.500000	641
10percentage	1500.0	78.384553	9.565983	43.0800	72.000000	
12graduation	1500.0	2008.122000	1.588542	2000.0000	2007.000000	2
12percentage	1500.0	74.947040	10.632432	40.8300	67.000000	
CollegeID	1500.0	5202.454667	4750.131676	2.0000	830.000000	3
CollegeTier	1500.0	1.926667	0.260770	1.0000	2.000000	
collegeGPA	1500.0	71.615147	8.747405	7.0000	67.147500	
CollegeCityID	1500.0	5202.454667	4750.131676	2.0000	830.000000	3
CollegeCityTier	1500.0	0.278000	0.448163	0.0000	0.000000	
GraduationYear	1500.0	2012.621333	1.288559	2008.0000	2012.000000	2
English	1500.0	501.883333	103.976105	195.0000	430.000000	
Logical	1500.0	501.210000	87.208808	215.0000	445.000000	
Quant	1500.0	514.505333	118.143311	120.0000	433.750000	
Domain	1500.0	0.509663	0.460305	-1.0000	0.352981	
ComputerProgramming	1500.0	357.774667	202.022875	-1.0000	305.000000	
ElectronicsAndSemicon	1500.0	93.853333	155.940551	-1.0000	-1.000000	
ComputerScience	1500.0	84.992000	171.721189	-1.0000	-1.000000	
MechanicalEngg	1500.0	22.992667	99.572364	-1.0000	-1.000000	
ElectricalEngg	1500.0	20.673333	98.467198	-1.0000	-1.000000	
TelecomEngg	1500.0	34.525333	106.704076	-1.0000	-1.000000	
CivilEngg	1500.0	3.997333	43.220335	-1.0000	-1.000000	
conscientiousness	1500.0	-0.038361	1.021743	-3.5085	-0.733500	
agreeableness	1500.0	0.183612	0.858094	-4.2831	-0.287100	
extraversion	1500.0	0.048418	0.919562	-3.3713	-0.598000	
nueroticism	1500.0	-0.091588	1.010601	-2.6430	-0.868200	
openess_to_experience	1500.0	-0.102384	0.908886	-5.8428	-0.669200	

1.3 Exploratory Data analysis

1.3.1 Visualization

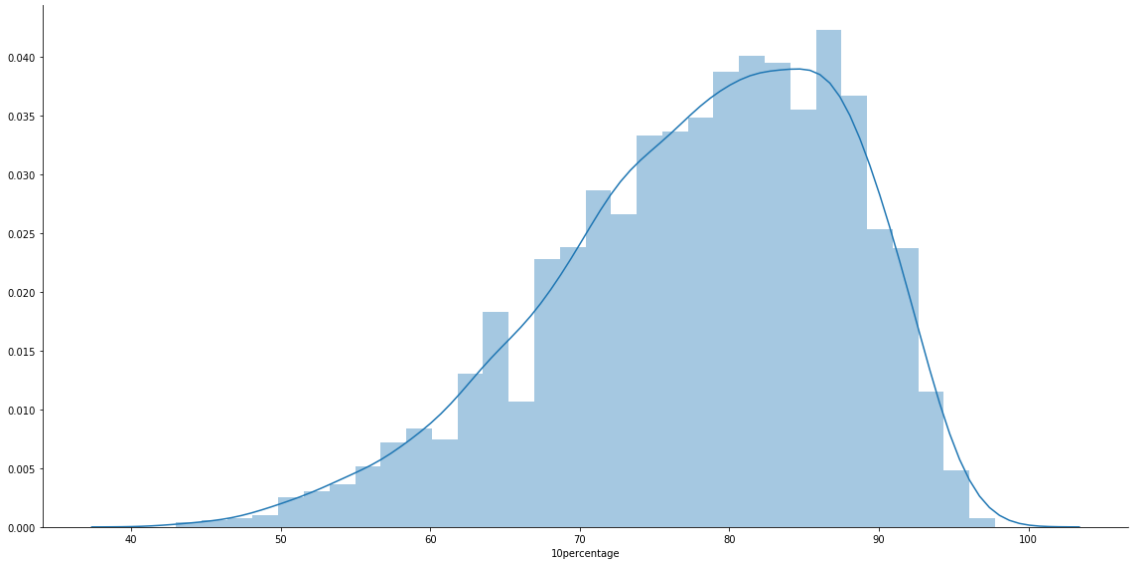
1.3.1.1 '10'nth percentage attribute

In [9]:

```
sns.FacetGrid(data=train,height=8,aspect=2).map(sns.distplot,"10percentage").add
_legend()
```

Out[9]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6cceb2a4a8>
```

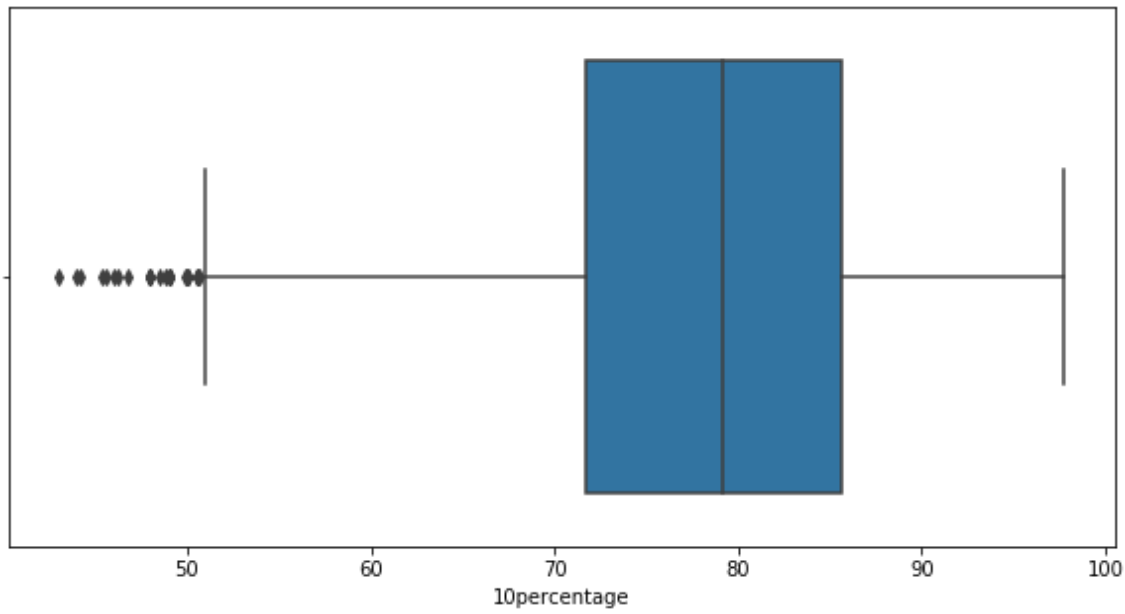


In [10]:

```
plt.figure(figsize=(10,5))
sns.boxplot(train['10percentage'])
```

Out[10]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6cced6f5c0>
```



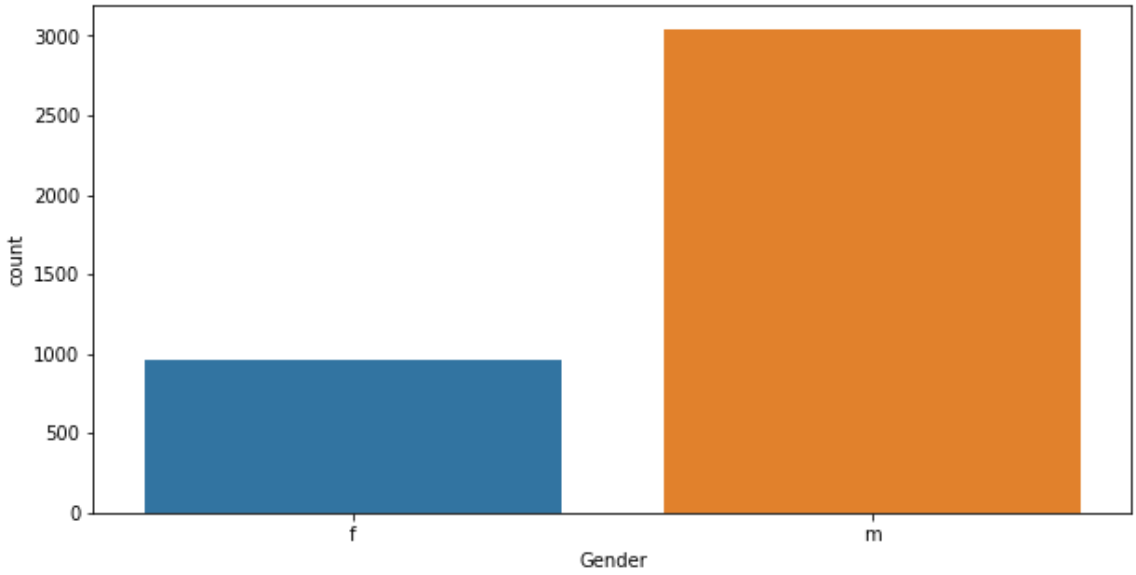
1.3.1.2 Gender attribute

In [11]:

```
plt.figure(figsize=(10,5))
sns.countplot(train['Gender'])
```

Out[11]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6ccf01fb70>
```



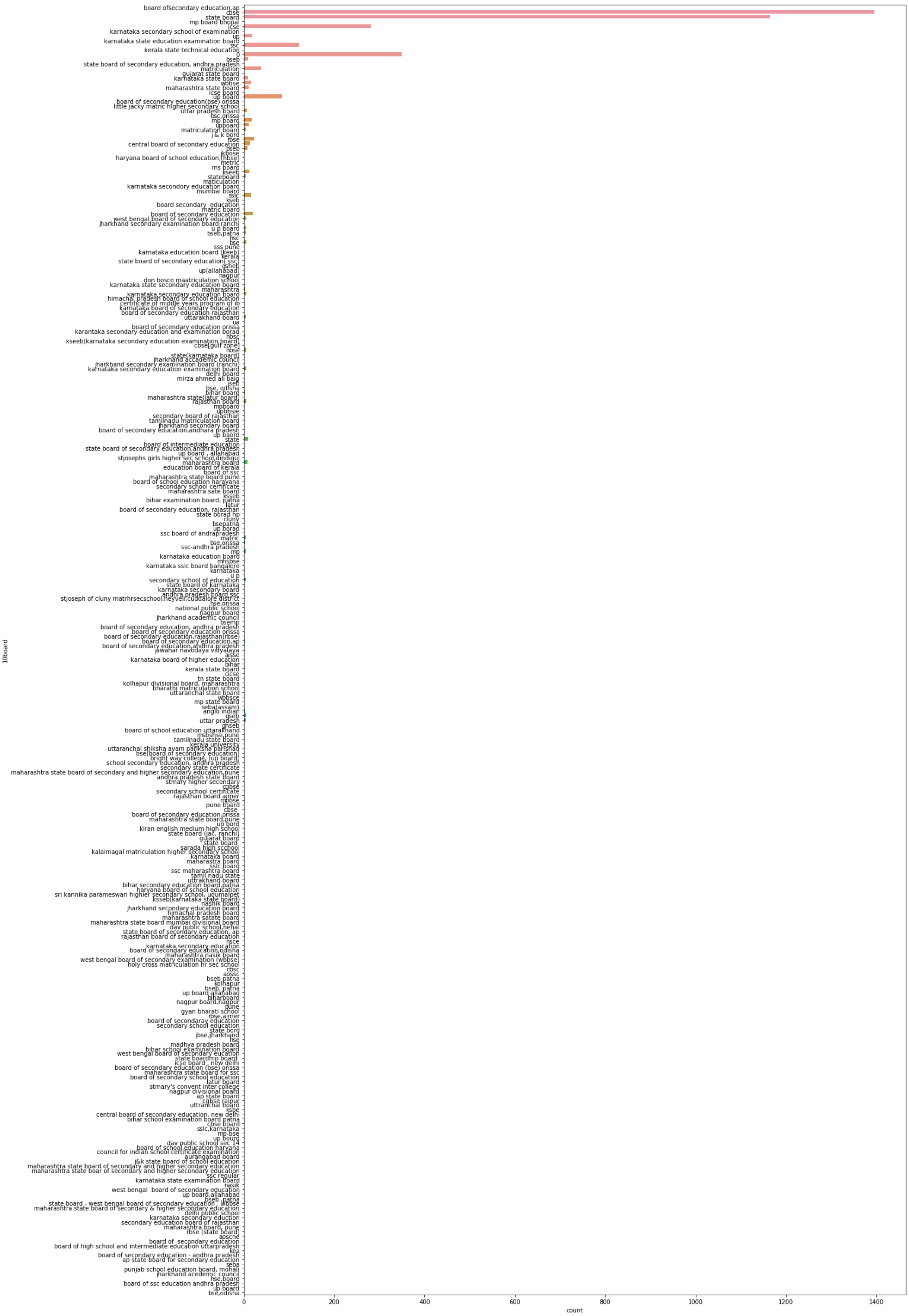
1.3.1.3 '10'nth Board attribute

In [12]:

```
plt.figure(figsize=(20,40))
sns.countplot(y=train['10board'])
```

Out[12]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6cceae0eb8>
```



10board
count
0
200
400
600
800
1000
1200
1400

In [13]:

```
train['10board'].unique()
```

Out[13]:

```
array(['board ofsecondary education,ap', 'cbse', 'state board',
       'mp board bhopal', 'icse',
       'karnataka secondary school of examination', 'up',
       'karnataka state education examination board', 'ssc',
       'kerala state technical education', 0, 'bseb',
       'state board of secondary education, andhra pradesh',
       'matriculation', 'gujarat state board', 'karnataka state boar
d',
       'wbbse', 'maharashtra state board', 'icse board', 'up board',
       'board of secondary education(bse) orissa',
       'little jacky matric higher secondary school',
       'uttar pradesh board', 'bsc,orissa', 'mp board', 'upboard',
       'matriculation board', 'j & k bord', 'rbse',
       'central board of secondary education', 'pseb', 'jkbose',
       'haryana board of school education,(hbse)', 'metric', 'ms boa
rd',
       'kseeb', 'stateboard', 'maticulation',
       'karnataka secondory education board', 'mumbai board', 'ssl
c',
       'kseb', 'board secondary  education', 'matric board',
       'board of secondary education',
       'west bengal board of secondary education',
       'jharkhand secondary examination board,ranchi', 'u p board',
       'bseb,patna', 'hsc', 'bse', 'sss pune',
       'karnataka education board (keeb)', 'kerala',
       'state board of secondary education( ssc)', 'gsheb',
       'up(allahabad)', 'nagpur', 'don bosco maatriculation school',
       'karnataka state secondary education board', 'maharashtra',
       'karnataka secondary education board',
       'himachal pradesh board of school education',
       'certificate of middle years program of ib',
       'karnataka board of secondary education',
       'board of secondary education rajasthan', 'uttarakhand boar
d',
       'ua', 'board of secendary education orissa',
       'karantaka secondary education and examination borad', 'hbs
c',
       'kseeb(karnataka secondary education examination board)',
       'cbse[gulf zone]', 'hbse', 'state(karnataka board)',
       'jharkhand accademic council',
       'jharkhand secondary examination board (ranchi)',
       'karnataka secondary education examination board', 'delhi boa
rd',
       'mirza ahmed ali baig', 'jseb', 'bse, odisha', 'bihar board',
       'maharashtra state(latur board)', 'rajasthan board', 'mpboar
d',
       'upbhsie', 'secondary board of rajasthan',
       'tamilnadu matriculation board', 'jharkhand secondary board',
       'board of secondary education,andhara pradesh', 'up baord',
       'state', 'board of intermediate education',
       'state board of secondary education,andhra pradesh',
       'up board , allahabad',
       'stjosephs girls higher sec school,dindigul', 'maharashtra bo
ard',
       'education board of kerala', 'board of ssc',
       'maharashtra state board pune',
       'board of school education harayana',
       'secondary school cerficate', 'maharashtra sate board', 'ks
seb',
```

```
       'bihar examination board, patna', 'latur',
       'board of secondary education, rajasthan', 'state borad hp',
       'cluny', 'bsepatna', 'up borad', 'ssc board of andrapradesh',
       'matric', 'bse,orissa', 'ssc-andhra pradesh', 'mp',
       'karnataka education board', 'mhsbse',
       'karnataka sslc board bangalore', 'karnataka', 'u p',
       'secondary school of education', 'state board of karnataka',
       'karnataka secondary board', 'andhra pradesh board ssc',
       'stjoseph of cluny matrhrsecschool,neyveli,cuddalore distric
t',
       'hse,orissa', 'national public school', 'nagpur board',
       'jharkhand academic council', 'bsemp',
       'board of secondary education, andhra pradesh',
       'board of secondary education orissa',
       'board of secondary education,rajasthan(rbse)',
       'board of secondary education,ap',
       'board of secondary education,andhra pradesh',
       'jawahar navodaya vidyalaya', 'aisse',
       'karnataka board of higher education', 'bihar',
       'kerala state board', 'cicse', 'tn state board',
       'kolhapur divisional board, maharashtra',
       'bharathi matriculation school', 'uttaranchal state board',
       'wbbsce', 'mp state board', 'seba(assam)', 'anglo indian', 'g
seb',
       'uttar pradesh', 'ghseb', 'board of school education uttarakh
and',
       'msbshse,pune', 'tamilnadu state board', 'kerala university',
       'uttaranchal shiksha avam pariksha parishad',
       'bse(board of secondary education)',
       'bright way college, (up board)',
       'school secondary education, andhra pradesh',
       'secondary state certificate',
       'maharashtra state board of secondary and higher secondary ed
ucation,pune',
       'andhra pradesh state board', 'stmary higher secondary', 'cgb
se',
       'secondary school certificate', 'rajasthan board ajmer', 'mpb
se',
       'pune board', 'cbse ', 'board of secondary education,orissa',
       'maharashtra state board,pune', 'up bord',
       'kiran english medium high school', 'state board (jac, ranch
i)',
       'gujarat board', 'state board ', 'sarada high scchool',
       'kalaimagal matriculation higher secondary school',
       'karnataka board', 'maharastra board', 'sslc board',
       'ssc maharashtra board', 'tamil nadu state', 'uttrakhand boar
d',
       'bihar secondary education board,patna',
       'haryana board of school education',
       'sri kannika parameswari highier secondary school, udumalpe
t',
       'ksseb(karnataka state board)', 'nashik board',
       'jharkhand secondary education board', 'himachal pradesh boar
d',
       'maharashtra satate board',
       'maharashtra state board mumbai divisional board',
       'dav public school,hehal',
       'state board of secondary education, ap',
       'rajasthan board of secondary education', 'hsce',
       'karnataka secondary education',
       'board of secondary education,odisha', 'maharashtra nasik boa
```

```
rd',
       'west bengal board of secondary examination (wbbse)',
       'holy cross matriculation hr sec school', 'cbsc', 'apssc',
       'bseb patna', 'kolhapur', 'bseb, patna', 'up board allahaba
d',
       'biharboard', 'nagpur board,nagpur', 'pune', 'gyan bharati sc
hool',
       'rbse,ajmer', 'board of secondaray education',
       'secondary school education', 'state bord', 'jbse,jharkhand',
       'hse', 'madhya pradesh board', 'bihar school examination boar
d',
       'west bengal board of secondary eucation', 'state boardmp boa
rd ',
       'icse board , new delhi',
       'board of secondary education (bse) orissa',
       'maharashtra state board for ssc',
       'board of secondary school education', 'latur board',
       "stmary's convent inter college", 'nagpur divisional board',
       'ap state board', 'cgbse raipur', 'uttranchal board', 'ksbe',
       'central board of secondary education, new delhi',
       'bihar school examination board patna', 'cbse board',
       'sslc,karnataka', 'mp-bse', 'up bourd', 'dav public school se
c 14',
       'board of school education haryana',
       'council for indian school certificate examination',
       'aurangabad board', 'j&k state board of school education',
       'maharashtra state board of secondary and higher secondary ed
ucation',
       'maharashtra state boar of secondary and higher secondary edu
cation',
       'ssc regular', 'karnataka state examination board', 'nasik',
       'west bengal  board of secondary education', 'up board,allaha
bad',
       'bseb ,patna',
       'state board - west bengal board of secondary education : wbb
se',
       'maharashtra state board of secondary & higher secondary educ
ation',
       'delhi public school', 'karnataka secondary eduction',
       'secondary education board of rajasthan',
       'maharashtra board, pune', 'rbse (state board)', 'apsche',
       'board of  secondary education',
       'board of high school and intermediate education uttarprades
h',
       'kea', 'board of secondary education - andhra pradesh',
       'ap state board for secondary education', 'seba',
       'punjab school education board, mohali',
       'jharkhand acedemic council', 'hse,board',
       'board of ssc education andhra pradesh', 'up-board', 'bse,odi
sha'],
      dtype=object)
```

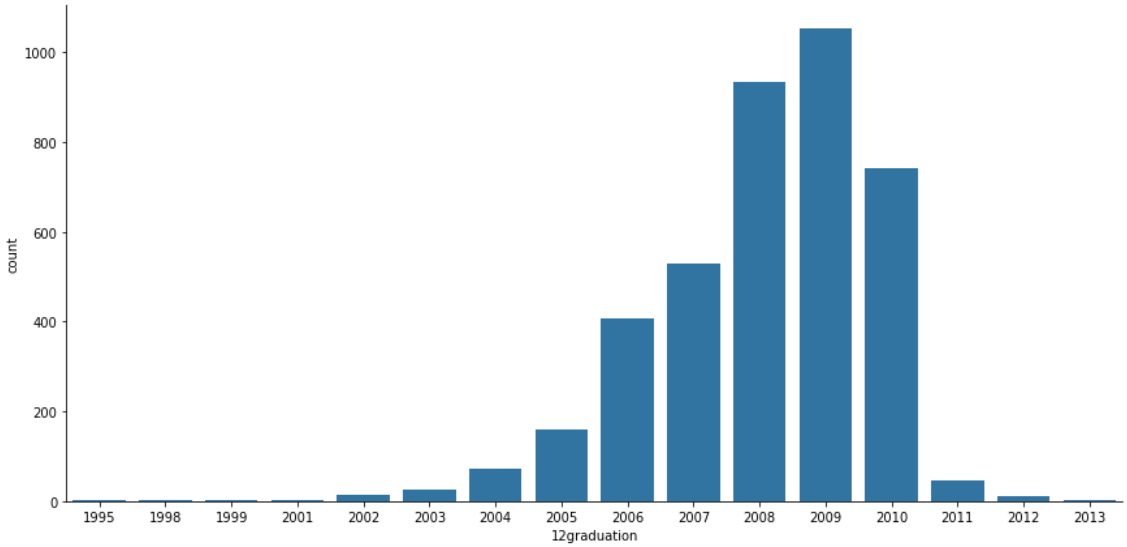
1.3.1.4 Year of 12th graduation attribute

In [14]:

```
sns.FacetGrid(data=train,height=6,aspect=2).map(sns.countplot,"12graduation").add_legend()
plt.ylabel("count")
```

Out[14]:

```
Text(7.197453703703708, 0.5, 'count')
```



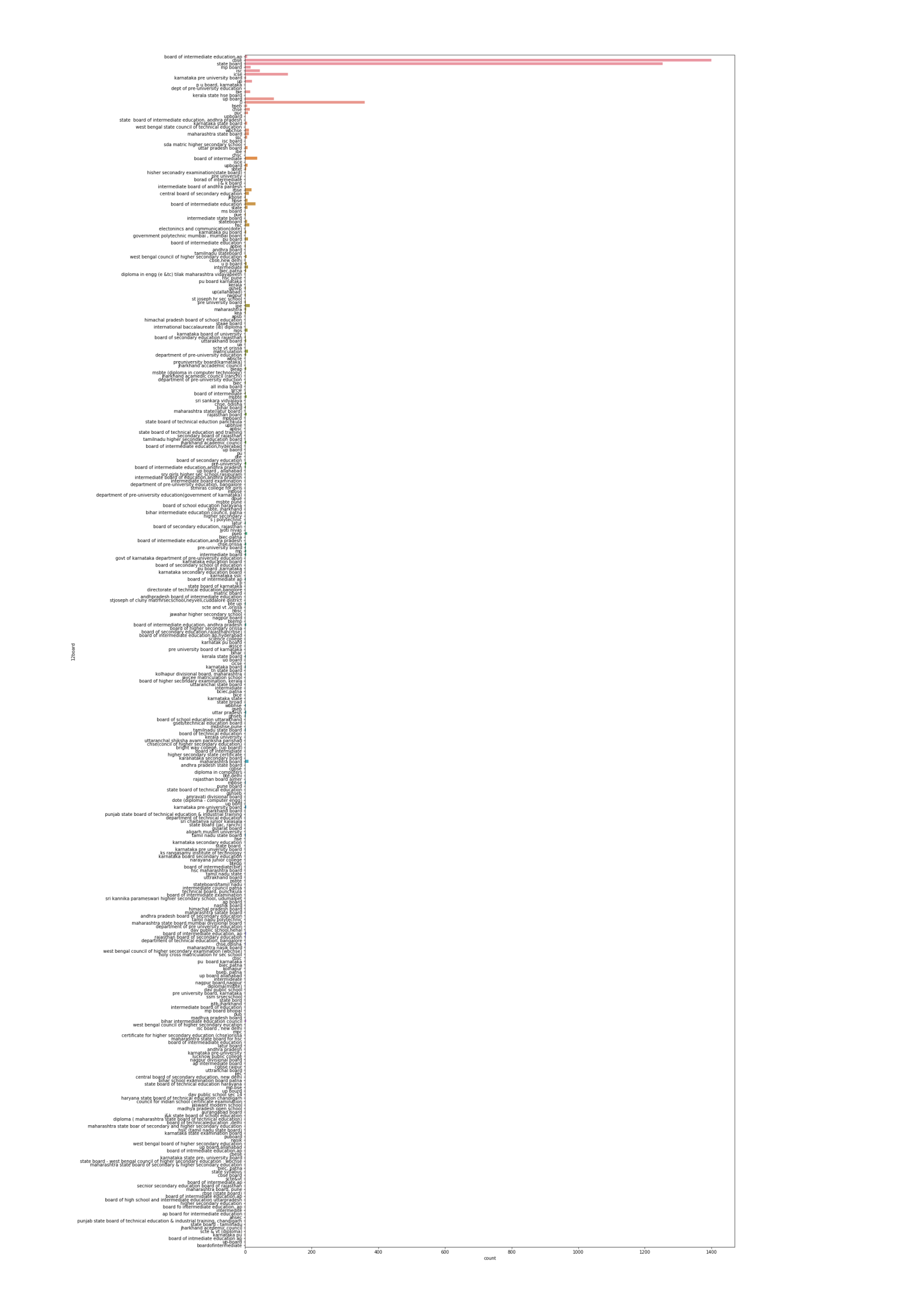
1.3.1.5 '12'th board attribute

In [15]:

```
plt.figure(figsize=(20,50))
sns.countplot(y=train['12board'])
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6cc0f8d470>
```


12board
count
0
200
400
600
800
1000
1200
1400
board of intermediate education,ap
cbse
state board
mp board
isc
icse
karnataka pre university board
p u board, karnataka
dept of pre-university education
bie
kerala state hse board
up board
bseb
ap
buc
upboard
state board of intermediate education, andhra pradesh
karnataka state board
west bengal state council of technical education
bseitse
maharashtra state board
isc
isc board
sda matric higher secondary school
uttar pradesh board
tbe
cbse
board of intermediate
icse
upboard
bseit
hisher seconadry examination(state board)
pre university
borad of intermediate
bsebbse
intermediate board of andhra pardesh
rbse
central board of secondary education
kbose
hbse
board of intermediate education
ap
ms board
pue
intermediate state board
stateboard
hsc
electroninc and communication(dote)
karnataka board
government polytechnic mumbai , mumbai board
pu board
baord of intermediate education
apbie
andhra board
tamilnadu stateboard
west bengal council of higher secondary education
cbse,new dedhi
u p board
interfediate
biec,patna
diploma in engg (e &tc) tilak maharashtra vidyapeeth
hsc pune
pu board karnataka
kerala
gshseb
up(allahabad)
nagpur
st joseph hr sec school
pre university board
maharashtra
bsea
apsbb
himachal pradesh board of school education
staee board
international baccalaureate (ib) diploma
hpbose
karnataka board of university
board of secondary education rajasthan
uttarakhand board
up
scte vt orissa
matriculation
department of pre-university education
bsebse
preuniversity board(karnataka)
jharkhand accademic council
bteap
msbte (diploma in computer technology)
jharkhand acamedic council (ranchi)
department of pre-university eduction
all india board
sircw
board of intermediate
mpbse
sri sankara vidyalaya
cbse, odisha
bihar board
maharashtra state(latur)
rajasthan board
mpboard
state board of technical eduction panchkula
upbhsie
apsbc
state board of technical education and training
secondary board of rajasthan
tamilnadu higher secondary education board
jharkhand academic council
board of intermediate education,hyderabad
up baord
jkbose
board of secondary education
pre-university
board of intermediate education,andhra pradesh
up board , allahabad
anu girls higher sec school,rasipuram
intermediate board of education,andhra pradesh
intermediate board examination
department of pre-university education
stmiras college for girls
mhbse
department of pre-university education(government of karnataka)
dbue
msbte pune
board of school education haryana
sbse, haryana
bihar intermediate education council, patna
higher secondary
s j polytechnic
latur
board of secondary education, rajasthan
jyoti nivas
bsebse
biec patna
board of intermediate education,andra pradesh
chse,orissa
pre-university board
hsc
intermediate board
govt of karnataka department of pre university education
karnataka education board
board of secondary school of education
pu board ,karnataka
karnataka secondary education board
karnataka sslc
board of intermediate ap
bsebse
state board of karnataka
directorate of technical education,banglore
matric board
andhpradesh board of intermediate education
stjoseph of cluny matrhrsecschool,neyveli,cuddalore district
bte up
scte and vt orissa
cbse
jawahar higher secondary school
nagpur board
bsemp
board of intermediate education, andhra pradesh
board of higher secondary orissa
board of secondary education, orissa
board of intermediate education,ap,hyderabad
science college
karnatak pu board
dsscce
pre university board of karnataka
bihar
kerala state board
up board
cicse
karnataka board
tn state board
kolhapur divisional board, maharashtra
aicyce matriculation schooi
board of higher secondary examination, kerala
uttaranchal state board
intermidiate
bciec,patna
ptce
karnataka state
state hmad
wbbsee
uttar pradesh
gtseb
board of school education uttarakhand
gseb(technical education board
mbbhse,pune
tamilnadu state board
board of technical education
kerala university
uttaranchal shiksha avam pariksha parishad
chse(concil of higher secondary education)
bright way college, (up board)
board of intermediate
higher secondary state certificate
karnataka secondary board
andhra pradesh state board
cgbse
diploma in computers
bse,delhi
rajasthan board ajmer
hpbose
pune board
state board of technical education
gshseb
amravati divisional board
dote (diploma - computer engg)
up bord
karnataka pre university board
cgbse raipur
punjab state board of technical education & industrial training
department of technical education
sri chaitanya junior kalasala
state board (jac, ranchi)
gujarat board
aligarh muslim university
tamil nadu state board
bse
karnataka secondary education
state board
karnataka pre university board
ks rangasamy institute of technology
karnataka board secondary education
narayana junior college
bsebse
board of intermediate(bie)
hsc maharashtra board
tamil nadu state
uttrakhand board
bsebe
stateboard/tamil nadu
intermediate council patna
technical board, punchkula
board of intermediate examination
sri kannika parameswari highier secondary school, udumalpet
ap board
nashik board
himachal pradesh board
maharashtra satate board
andhra pradesh board of secondary education
tamil nadu polytechnic
maharashtra state board,mumbai divisional board
department of pre university education
dav public school,nehai
board of intermediate education
rajasthan board of secondary education
department of technical education, bangalore
chse,odisha
maharashtra nasik board
west bengal council of higher secondary examination (wbchse)
holy cross matriculation hr sec school
bsebe
pu board karnataka
biec patna
kolhapur
bseb, patna
up board allahabad
intermidieate
nagpur board,mumbai
dpsmm,madurai
dav public school
pre university board, karnataka
ssm srsecschool
state bord
stb,jharkhand
intermediate board of education
mp board bhopal
bsebse
madhya pradesh board
bihar intermediate education council
west bengal council of higher secondary eucation
isc board , new delhi
mpc
certificate for higher secondary education (chse),orissa
maharashtra state board for hse
board of intermediate education
latur board
andhra pradesh
karnataka pre-university
lucknow public college
nagpur divisional board
ap intermediate board
cgbse raipur
uttranchal board
hsc
central board of secondary education, new delhi
bihar school examination board patna
state board of technical education haryana
bsebse
up board
dav public school sec 14
haryana state board of technical education chandigarh
council for indian school certificate examination
jawaharl modern school
madhya pradesh board
aurangabad board
ijkk state board of school education
diploma (maharashtra state board of technical education)
board of technicaleducation, delhi
maharashtra state boar of secondary and higher secondary education
hsc (tamil nadu state board)
karnataka state examination board
pu board
nasik
west bengal board of higher secondary education
up board,allahabad
board of intrmediate education,ap
bseb
karnataka state pre- university board
state board , west bengal council of higher secondary education
maharashtra state board of secondary & higher secondary education
biec, patna
state syllabus
cbse board
scte&vt
board of intermediate,ap
secnior secondary education board of rajasthan
maharashtra board, pune
hbse (state board)
board of intermediate education,ap
board of high school and intermediate education uttarpradesh
higher secondary education
board fo intermediate education, ap
bsebse
ap board for intermediate education
ahsec
punjab state board of technical education & industrial training, chandigarh
state board - tamilnadu
jharkhand academic council
scte & vt (diploma)
bse, odisha
board of intmediate education ap
up-board
boardofintermediate

In [16]:

```
print(train['12board'].unique())
```

```
['board of intermediate education,ap' 'cbse' 'state board' 'mp boar
d'
 'isc' 'icse' 'karnataka pre university board' 'up' 'p u board, karn
ataka'
 'dept of pre-university education' 'bie' 'kerala state hse board'
 'up board' 0 'bseb' 'chse' 'puc' ' upboard'
 'state  board of intermediate education, andhra pradesh'
 'karnataka state board'
 'west bengal state council of technical education' 'wbchse'
 'maharashtra state board' 'ssc' 'isc board'
 'sda matric higher secondary school' 'uttar pradesh board' 'ibe' 'c
hsc'
 'board of intermediate' 'isce' 'upboard' 'sbtet'
 'hisher seconadry examination(state board)' 'pre university'
 'borad of intermediate' 'j & k board'
 'intermediate board of andhra pardesh' 'rbse'
 'central board of secondary education' 'jkbose' 'hbse'
 'board of intermediate education' 'state' 'ms board' 'pue'
 'intermediate state board' 'stateboard' 'hsc'
 'electonincs and communication(dote)' 'karnataka pu board'
 'government polytechnic mumbai , mumbai board' 'pu board'
 'baord of intermediate education' 'apbie' 'andhra board'
 'tamilnadu stateboard'
 'west bengal council of higher secondary education' 'cbse,new delh
i'
 'u p board' 'intermediate' 'biec,patna'
 'diploma in engg (e &tc) tilak maharashtra vidayapeeth' 'hsc pune'
 'pu board karnataka' 'kerala' 'gsheb' 'up(allahabad)' 'nagpur'
 'st joseph hr sec school' 'pre university board' 'ipe' 'maharashtr
a'
 'kea' 'apsb' 'himachal pradesh board of school education' 'staae bo
ard'
 'international baccalaureate (ib) diploma' 'nios'
 'karnataka board of university' 'board of secondary education rajas
than'
 'uttarakhand board' 'ua' 'scte vt orissa' 'matriculation'
 'department of pre-university education' 'wbscte'
 'preuniversity board(karnataka)' 'jharkhand accademic council' 'bie
ap'
 'msbte (diploma in computer technology)'
 'jharkhand acamedic council (ranchi)'
 'department of pre-university eduction' 'biec' 'all india board' 's
jrcw'
 ' board of intermediate' 'msbte' 'sri sankara vidyalaya' 'chse, odi
sha'
 'bihar board' 'maharashtra state(latur board)' 'rajasthan board'
 'mpboard' 'state board of technical eduction panchkula' 'upbhsie'
'apbsc'
 'state board of technical education and training'
 'secondary board of rajasthan'
 'tamilnadu higher secondary education board' 'jharkhand academic co
uncil'
 'board of intermediate education,hyderabad' 'up baord' 'pu' 'dte'
 'board of secondary education' 'pre-university'
 'board of intermediate education,andhra pradesh' 'up board , allaha
bad'
 'srv girls higher sec school,rasipuram'
 'intermediate board of education,andhra pradesh'
 'intermediate board examination'
 'department of pre-university education, bangalore'
 'stmiras college for girls' 'mbose'
```

```
 'department of pre-university education(government of karnataka)'
'dpue'
 'msbte pune' 'board of school education harayana' 'sbte, jharkhand'
 'bihar intermediate education council, patna' 'higher secondary'
 's j polytechnic' 'latur' 'board of secondary education, rajasthan'
 'jyoti nivas' 'pseb' 'biec-patna'
 'board of intermediate education,andra pradesh' 'chse,orissa'
 'pre-university board' 'mp' 'intermediate board'
 'govt of karnataka department of pre-university education'
 'karnataka education board' 'board of secondary school of educatio
n'
 'pu board ,karnataka' 'karnataka secondary education board'
 'karnataka sslc' 'board of intermediate ap' 'u p'
 'state board of karnataka' 'directorate of technical education,bang
lore'
 'matric board' 'andhpradesh board of intermediate education'
 'stjoseph of cluny matrhrsecschool,neyveli,cuddalore district' 'bte
up'
 'scte and vt ,orissa' 'hbsc' 'jawahar higher secondary school'
 'nagpur board' 'bsemp' 'board of intermediate education, andhra pra
desh'
 'board of higher secondary orissa'
 'board of secondary education,rajasthan(rbse)'
 'board of intermediate education:ap,hyderabad' 'science college'
 'karnatak pu board' 'aissce' 'pre university board of karnataka' 'b
ihar'
 'kerala state board' 'uo board' 'cicse' 'karnataka board'
 'tn state board' 'kolhapur divisional board, maharashtra'
 'jaycee matriculation school'
 'board of higher secondary examination, kerala' 'uttaranchal state
board'
 'intermidiate' 'bciec,patna' 'bice' 'karnataka state' 'state broad'
 'wbbhse' 'gseb' 'uttar pradesh' 'ghseb'
 'board of school education uttarakhand' 'gseb/technical education b
oard'
 'msbshse,pune' 'tamilnadu state board' 'board of technical educatio
n'
 'kerala university' 'uttaranchal shiksha avam pariksha parishad'
 'chse(concil of higher secondary education)'
 'bright way college, (up board)' 'board of intermidiate'
 'higher secondary state certificate' 'karanataka secondary board'
 'maharashtra board' 'andhra pradesh state board' 'cgbse'
 'diploma in computers' 'bte,delhi' 'rajasthan board ajmer' 'mpbse'
 'pune board' 'state board of technical education' 'gshseb'
 'amravati divisional board' 'dote (diploma - computer engg)' 'up bo
rd'
 'karnataka pre-university board' 'jharkhand board'
 'punjab state board of technical education & industrial training'
 'department of technical education' 'sri chaitanya junior kalasala'
 'state board (jac, ranchi)' 'gujarat board' 'aligarh muslim univers
ity'
 'tamil nadu state board' 'hse' 'karnataka secondary education'
 'state board ' 'karnataka pre unversity board'
 'ks rangasamy institute of technology'
 'karnataka board secondary education' 'narayana junior college' 'bt
eup'
 'board of intermediate(bie)' 'hsc maharashtra board' 'tamil nadu st
ate'
 'uttrakhand board' 'psbte' 'stateboard/tamil nadu'
 'intermediate council patna' 'technical board, punchkula'
 'board of intermidiate examination'
```

```
 'sri kannika parameswari highier secondary school, udumalpet' 'ap b
oard'
 'nashik board' 'himachal pradesh board' 'maharashtra satate board'
 'andhra pradesh board of secondary education' 'tamil nadu polytechn
ic'
 'maharashtra state board mumbai divisional board'
 'department of pre university education' 'dav public school,hehal'
 'board of intermediate education, ap'
 'rajasthan board of secondary education'
 'department of technical education, bangalore' 'chse,odisha'
 'maharashtra nasik board'
 'west bengal council of higher secondary examination (wbchse)'
 'holy cross matriculation hr sec school' 'cbsc' 'pu  board karnatak
a'
 'biec patna' 'kolhapur' 'bseb, patna' 'up board allahabad' 'intermi
deate'
 'nagpur board,nagpur' 'diploma(msbte)' 'dav public school'
 'pre university board, karnataka' 'ssm srsecschool' 'state bord'
 'jstb,jharkhand' 'intermediate board of education' 'mp board bhopa
l'
 'pub' 'madhya pradesh board' 'bihar intermediate education council'
 'west bengal council of higher secondary eucation'
 'isc board , new delhi' 'mpc'
 'certificate for higher secondary education (chse)orissa'
 'maharashtra state board for hsc' 'board of intermeadiate educatio
n'
 'latur board' 'andhra pradesh' 'karnataka pre-university'
 'lucknow public college' 'nagpur divisional board'
 'ap intermediate board' 'cgbse raipur' 'uttranchal board' 'jiec'
 'central board of secondary education, new delhi'
 'bihar school examination board patna'
 'state board of technical education harayana' 'mp-bse' 'up bourd'
 'dav public school sec 14'
 'haryana state board of technical education chandigarh'
 'council for indian school certificate examination'
 'jaswant modern school' 'madhya pradesh open school' 'aurangabad bo
ard'
 'j&k state board of school education'
 'diploma ( maharashtra state board of technical education)'
 'board of technicaleducation ,delhi'
 'maharashtra state boar of secondary and higher secondary educatio
n'
 'hslc (tamil nadu state board)' 'karnataka state examination board'
 'puboard' 'nasik' 'west bengal board of higher secondary education'
 'up board,allahabad' 'board of intrmediate education,ap' 'cbese'
 'karnataka state pre- university board'
 'state board - west bengal council of higher secondary education :
wbchse'
 'maharashtra state board of secondary & higher secondary education'
 'biec, patna' 'state syllabus' 'cbse board' 'scte&vt'
 'board of intermediate,ap'
 'secnior secondary education board of rajasthan'
 'maharashtra board, pune' 'rbse (state board)'
 'board of intermidiate education,ap'
 'board of high school and intermediate education uttarpradesh'
 'higher secondary education' 'board fo intermediate education, ap'
 'intermedite' 'ap board for intermediate education' 'ahsec'
 'punjab state board of technical education & industrial training, c
handigarh'
 'state board - tamilnadu' 'jharkhand acedemic council'
 'scte & vt (diploma)' 'karnataka pu' 'board of intmediate education
```

```
ap'
 'up-board' 'boardofintermediate']
```

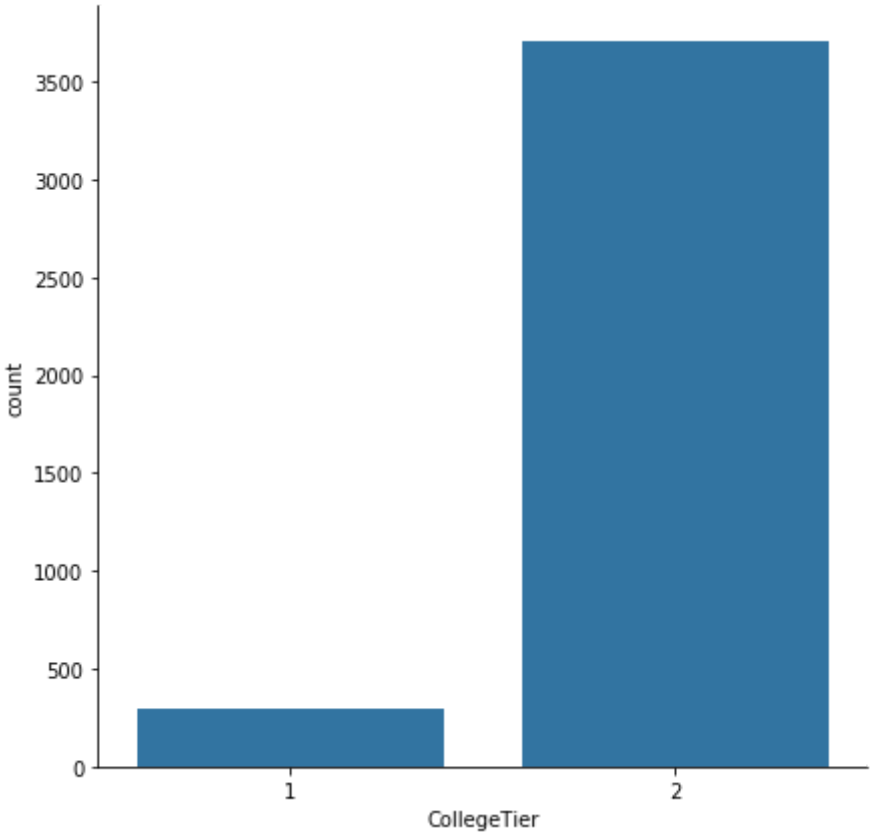
1.3.1.6 College tier

In [17]:

```
sns.FacetGrid(data=train,height=6).map(sns.countplot,"CollegeTier").add_legend()
plt.ylabel("count")
```

Out[17]:

```
Text(7.5972222222222285, 0.5, 'count')
```



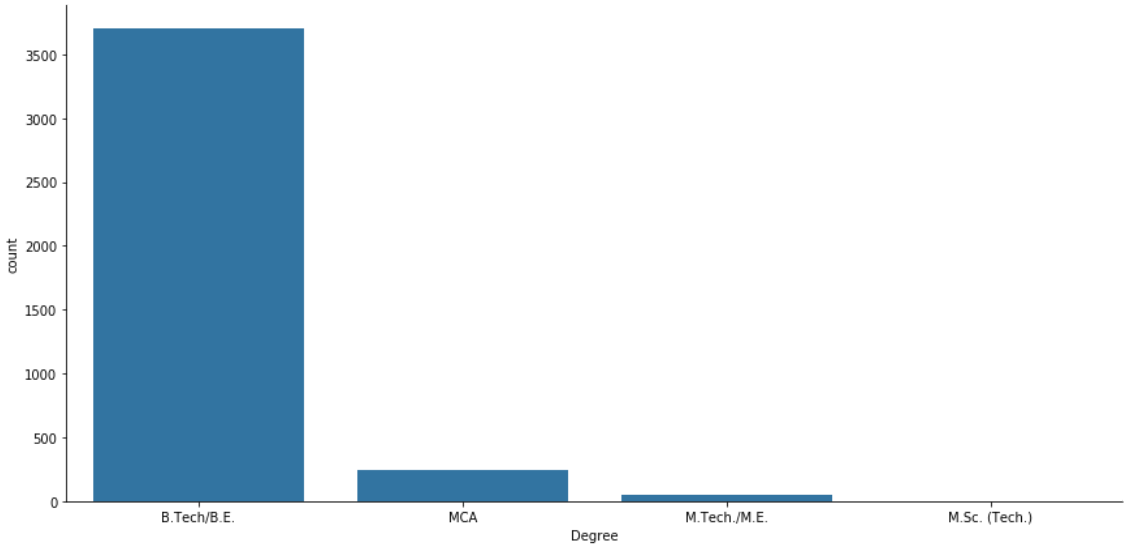
1.3.1.7 Degree attribute

In [18]:

```
sns.FacetGrid(data=train,height=6,aspect=2).map(sns.countplot,"Degree").add_lege
nd()
plt.ylabel("count")
```

Out[18]:

```
Text(7.198611111111113, 0.5, 'count')
```



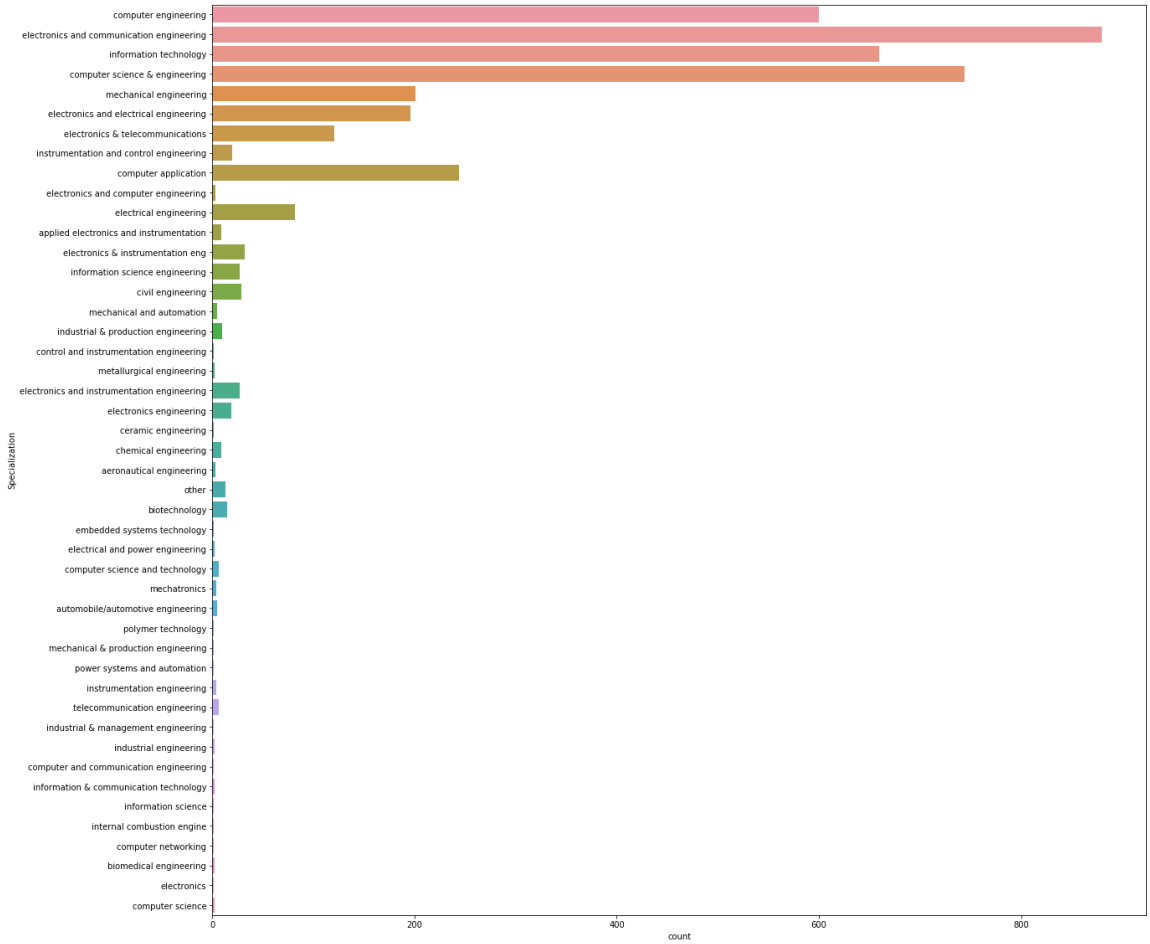
1.3.1.8 Specalization attribute

In [19]:

```
plt.figure(figsize=(20,20))
sns.countplot(y=train['Specialization'])
```

Out[19]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6cbfdad400>
```



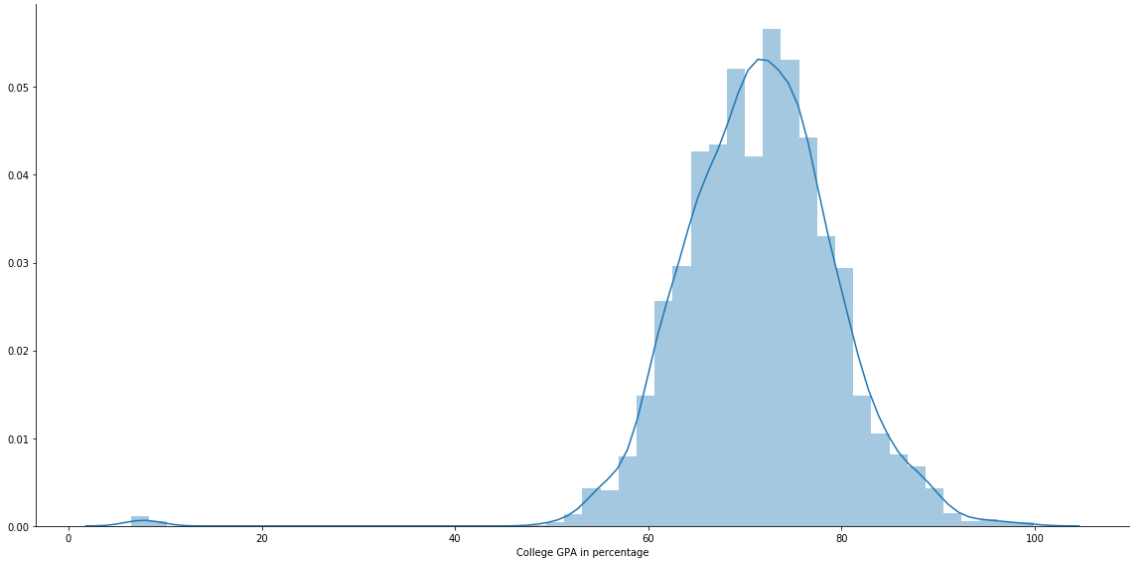
1.3.1.9 Graduation GPA (in percentage) attribute

In [20]:

```
sns.FacetGrid(data=train,height=8,aspect=2).map(sns.distplot,"collegeGPA").add_l
egend()
plt.xlabel("College GPA in percentage")
```

Out[20]:

```
Text(0.5, 20.80000000000001, 'College GPA in percentage')
```



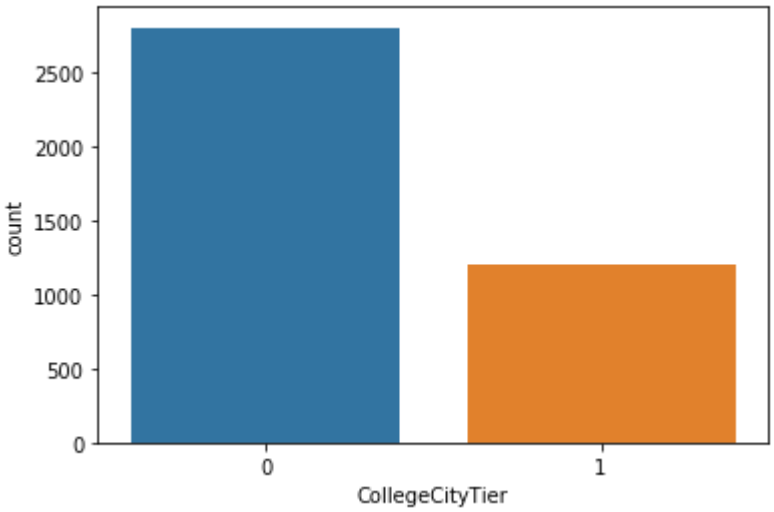
1.3.1.10 College City Tier attribute

In [21]:

```
sns.countplot(train['CollegeCityTier'])
```

Out[21]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6cbdda64a8>
```



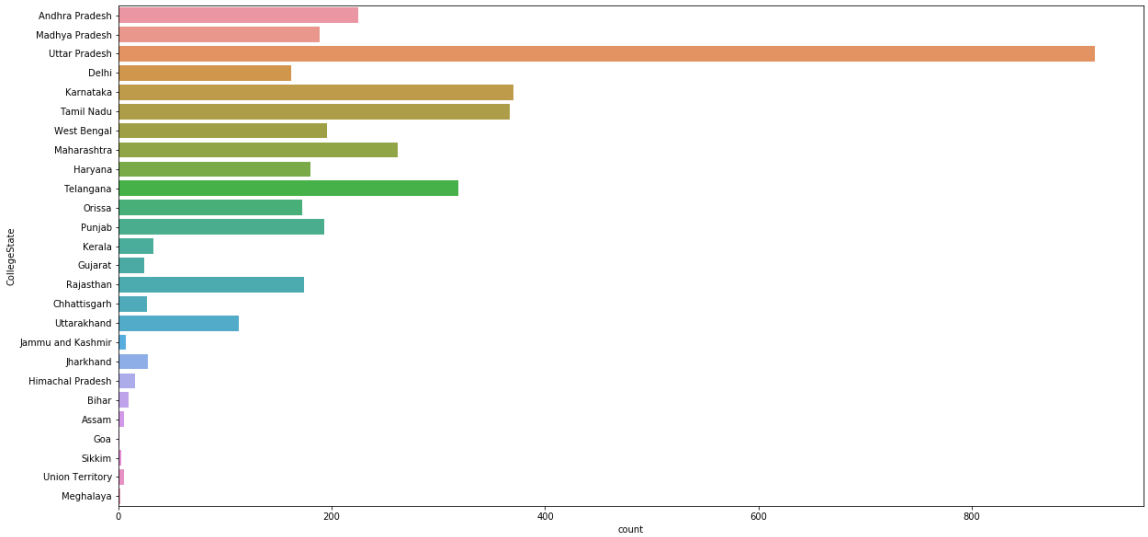
1.3.1.11 College state

In [22]:

```
plt.figure(figsize=(20,10))
sns.countplot(y=train['CollegeState'])
```

Out[22]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6cbdd6b780>
```



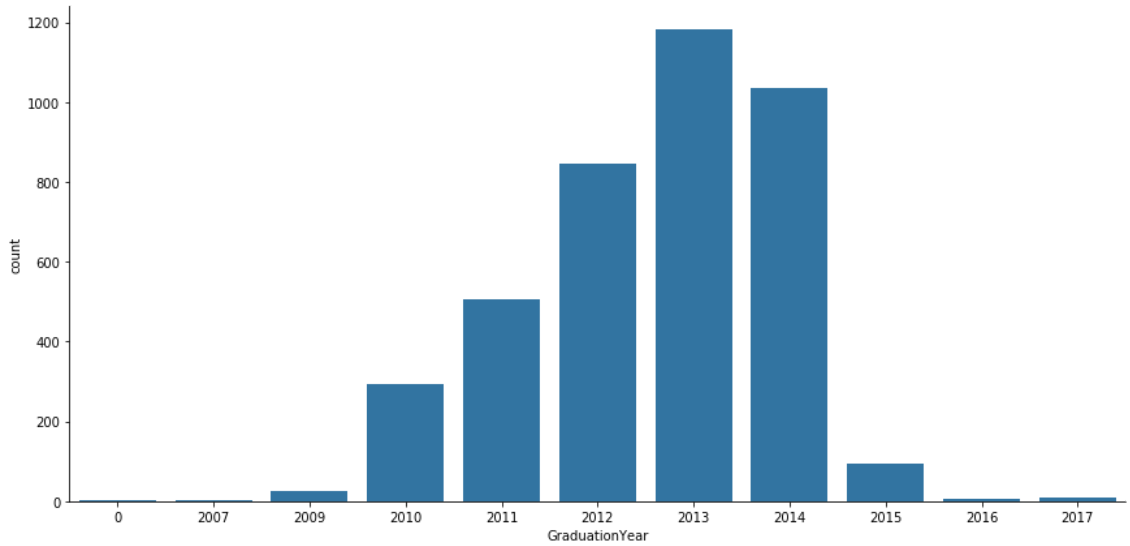
1.3.1.12 Year of bachelors degree graduation

In [23]:

```
sns.FacetGrid(data=train,height=6,aspect=2).map(sns.countplot,"GraduationYear").
add_legend()
plt.ylabel("count")
```

Out[23]:

```
Text(7.198611111111113, 0.5, 'count')
```



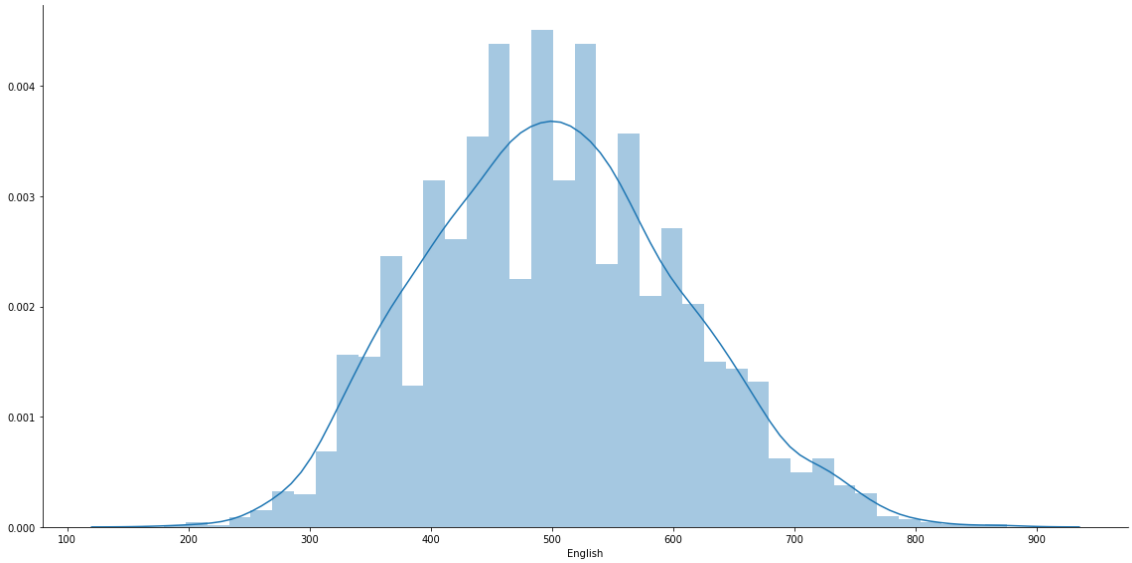
1.3.1.13 Amcat English section score attribute

In [24]:

```
sns.FacetGrid(data=train,height=8,aspect=2).map(sns.distplot,"English").add_legend()
```

Out[24]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6ccf01b898>
```



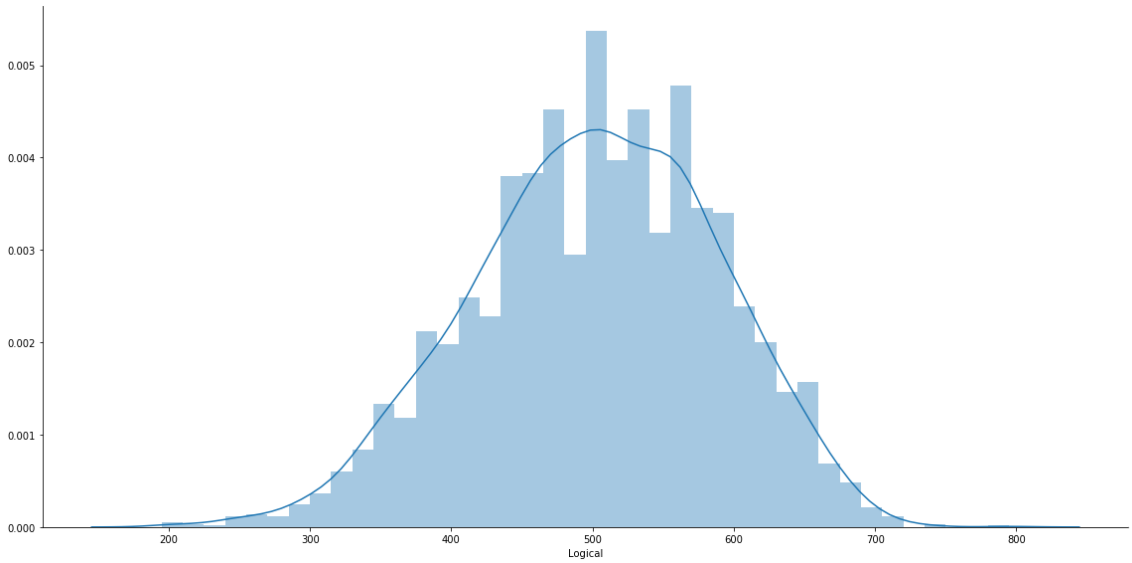
1.3.1.14 Amcat Logical ability section score attribute

In [25]:

```
sns.FacetGrid(data=train,height=8,aspect=2).map(sns.distplot,"Logical").add_legend()
```

Out[25]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6cbdab9908>
```



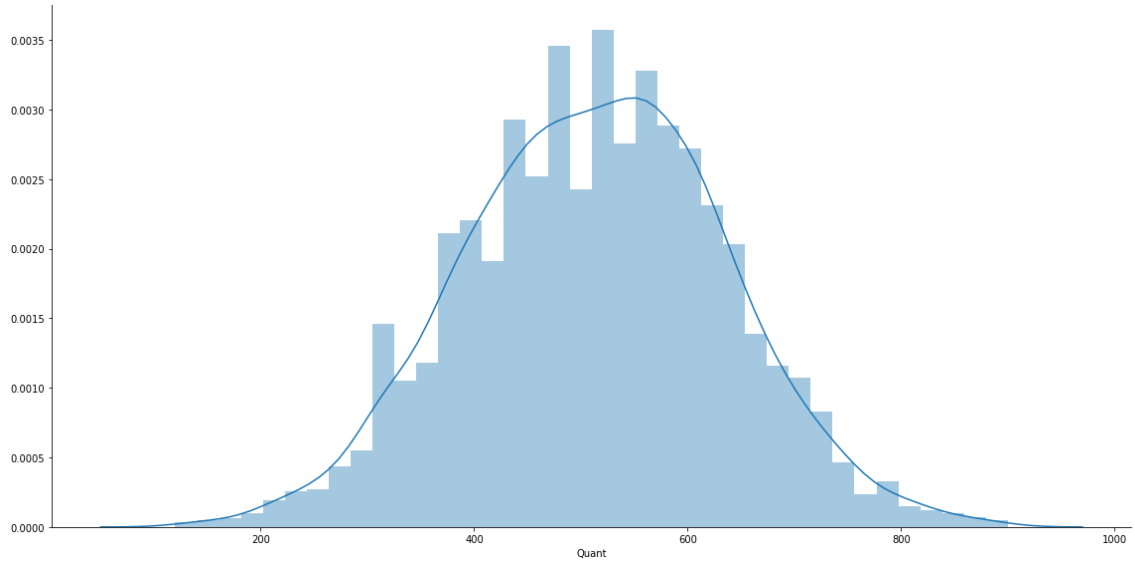
1.3.1.15 Amcat Quants section score attribute

In [26]:

```
sns.FacetGrid(data=train,height=8,aspect=2).map(sns.distplot,"Quant").add_legend
()
```

Out[26]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6cbda5ec88>
```



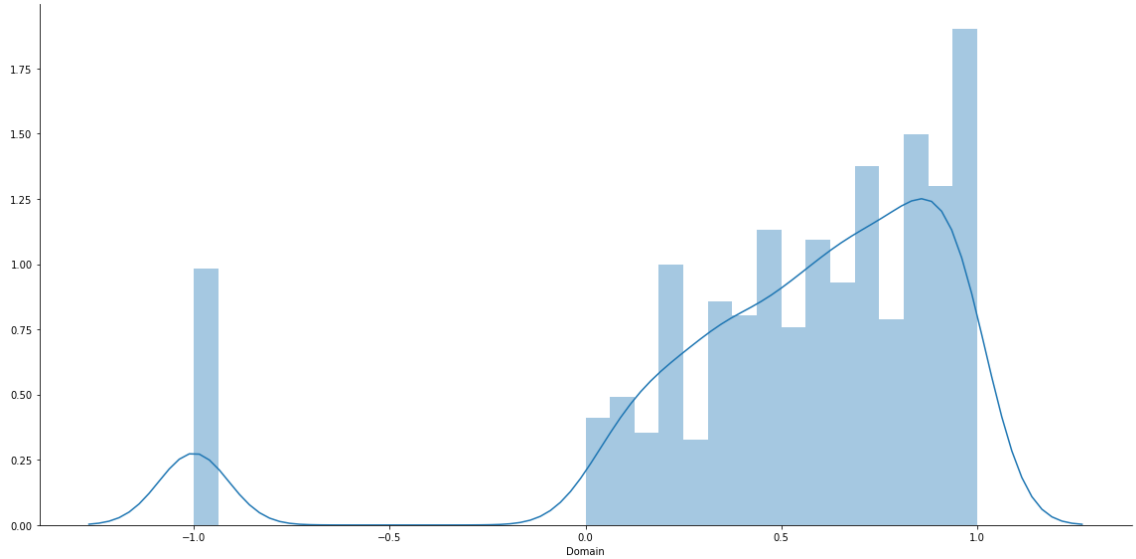
1.3.1.16 Amcat domain score

In [27]:

```
sns.FacetGrid(data=train,height=8,aspect=2).map(sns.distplot,"Domain").add_legen
d()
```

Out[27]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6cbda737b8>
```



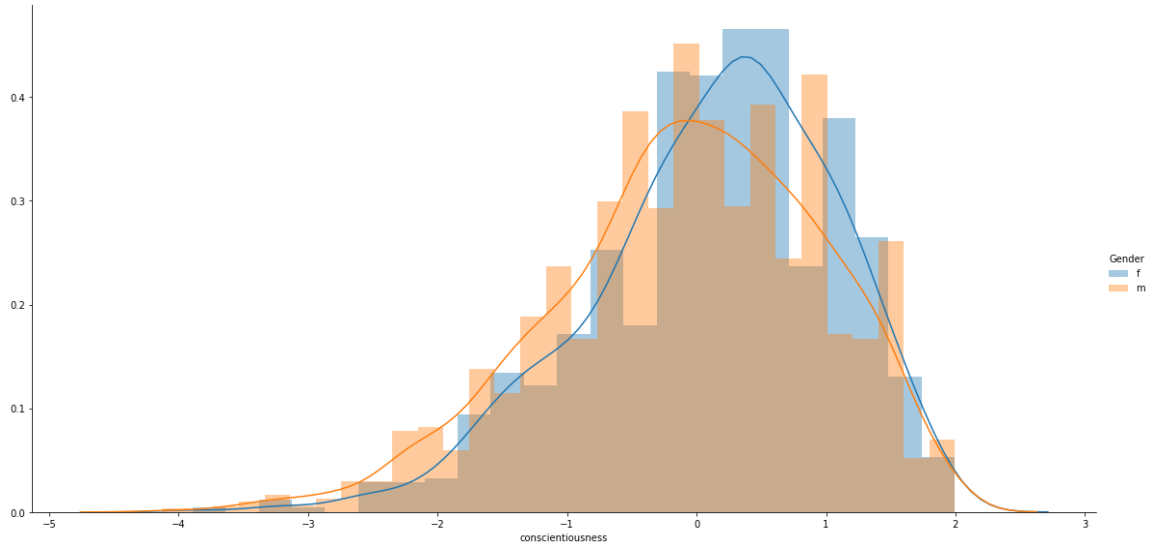
1.3.1.17 Personality test conscientiousness score attribute

In [28]:

```
sns.FacetGrid(data=train,hue="Gender",height=8,aspect=2).map(sns.distplot,"conscientiousness").add_legend()
```

Out[28]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6cbd8f89b0>
```



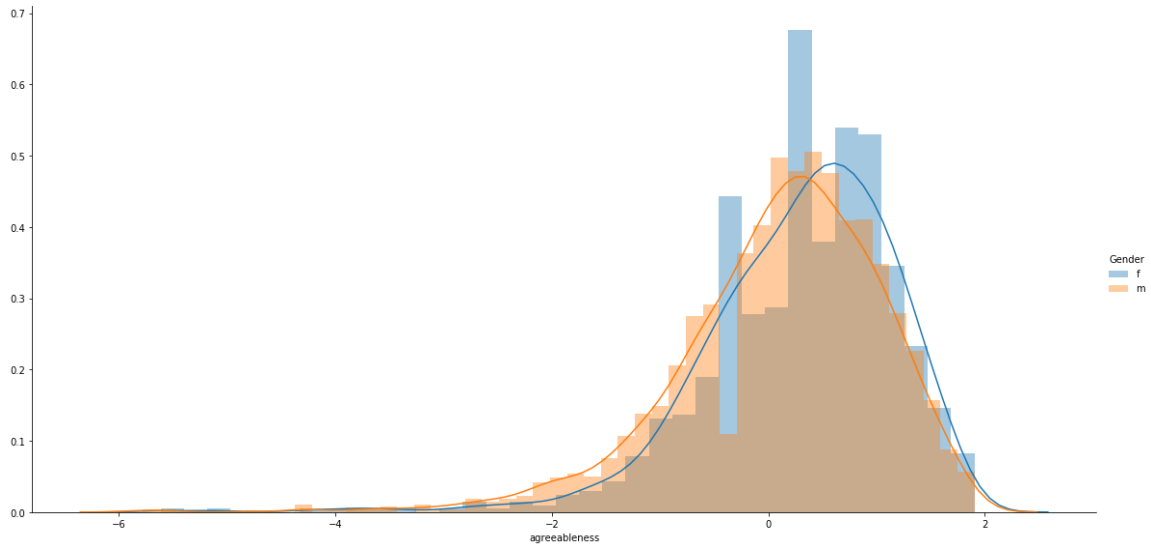
1.3.1.18 Personality test agreeableness score attribute

In [29]:

```
sns.FacetGrid(data=train,hue="Gender",height=8,aspect=2).map(sns.distplot,"agreeableness").add_legend()
```

Out[29]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6cbd8d8a20>
```



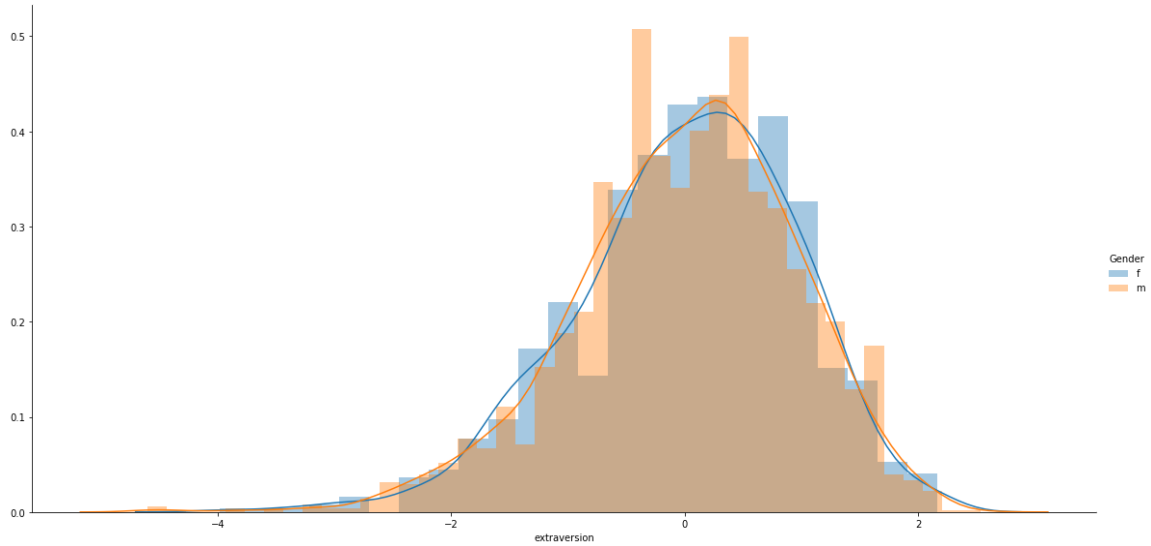
1.3.1.19 Personality test extraversion score attribute

In [30]:

```
sns.FacetGrid(data=train,hue="Gender",height=8,aspect=2).map(sns.distplot,"extra
version").add_legend()
```

Out[30]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6cbd7dac88>
```



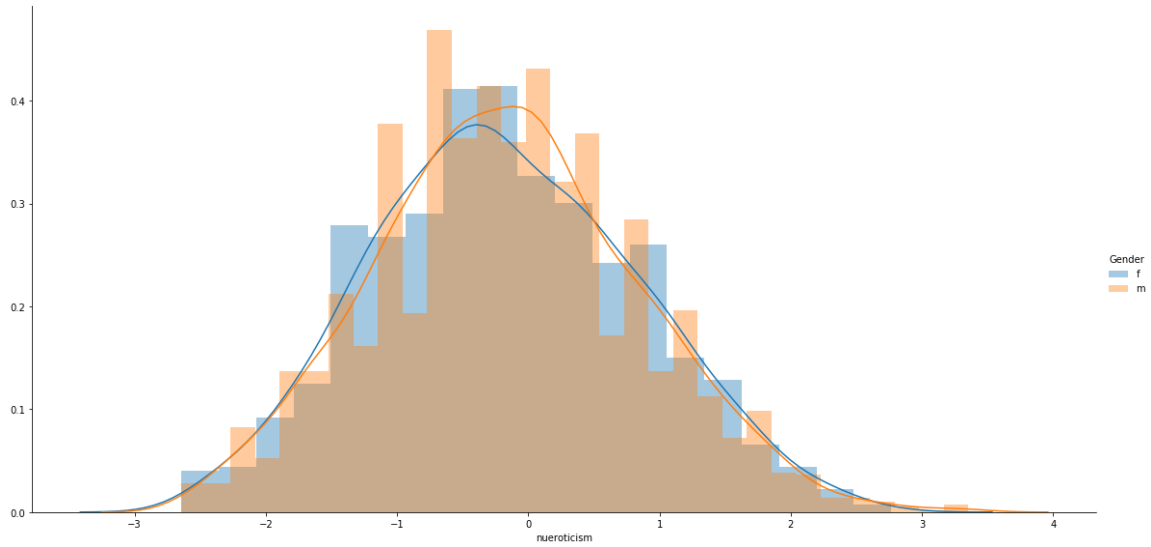
1.3.1.20 Personality test nueroticism score attribute

In [31]:

```
sns.FacetGrid(data=train,hue="Gender",height=8,aspect=2).map(sns.distplot,"nuero
ticism").add_legend()
```

Out[31]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6cbd5e0630>
```



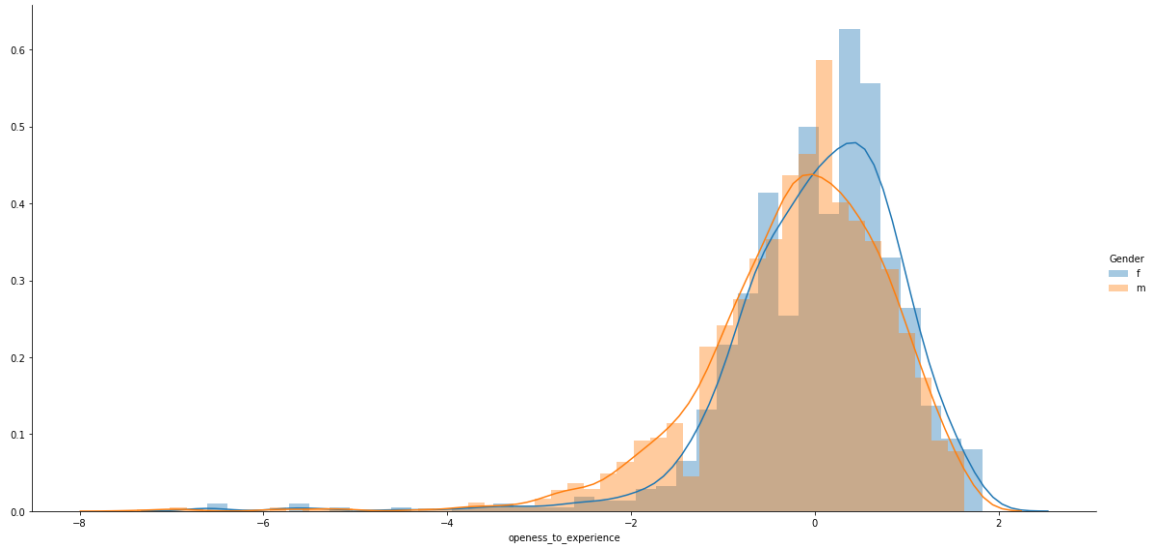
1.3.1.21 Personality test openess_to_experience score attribute

In [32]:

```
sns.FacetGrid(data=train,hue="Gender",height=8,aspect=2).map(sns.distplot,"openess_to_experience").add_legend()
```

Out[32]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6cbd45bfd0>
```



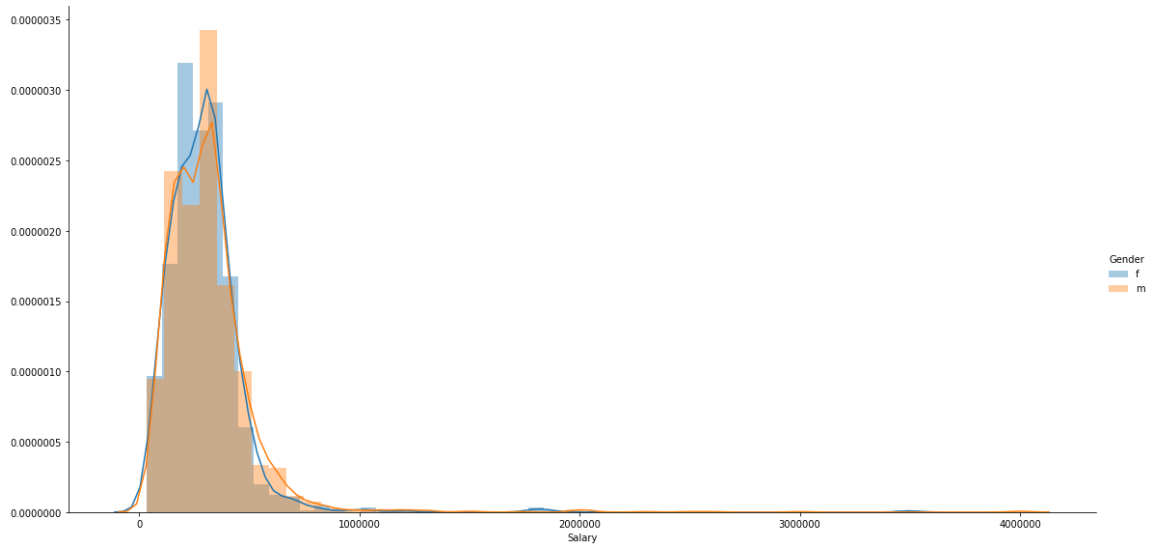
1.3.1.22 Analyzing the target (Salary) Variable by gender

In [33]:

```
sns.FacetGrid(data=train,hue="Gender",height=8,aspect=2).map(sns.distplot,"Salary").add_legend()
```

Out[33]:

```
<seaborn.axisgrid.FacetGrid at 0x7f6cbd308710>
```



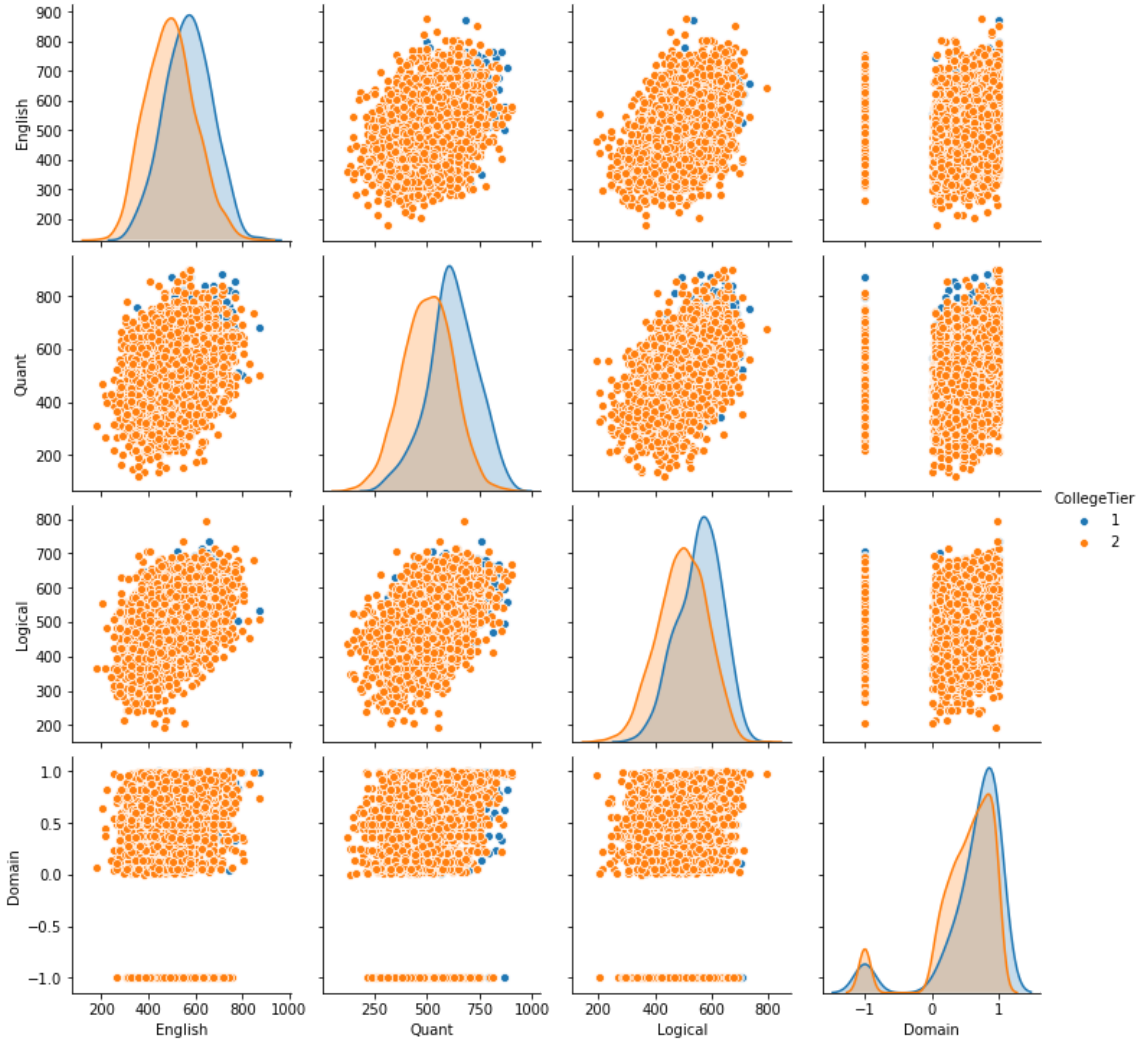
1.3.1.23 Pair plot of amcat scores by College Tier

In [34]:

```
amcatScores = train[['English','Quant','Logical','Domain','CollegeTier']]
sns.pairplot(data=amcatScores,hue="CollegeTier")
```

Out[34]:

```
<seaborn.axisgrid.PairGrid at 0x7f6cbd308080>
```



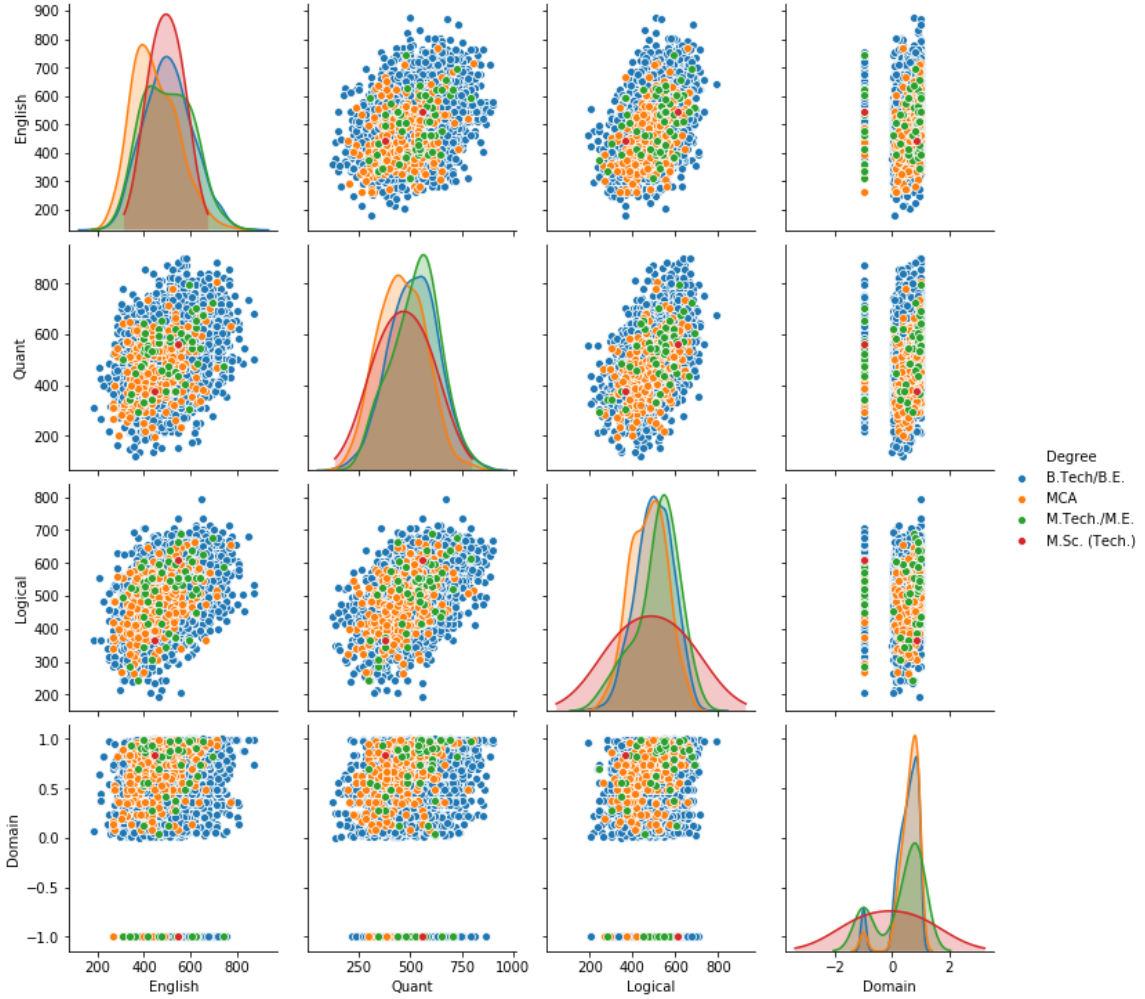
1.3.1.24 Pair plot of amcat scores by Degree

In [35]:

```
amcatScores1 = train[['English','Quant','Logical','Domain','Degree']]
sns.pairplot(data=amcatScores1,hue="Degree")
```

Out[35]:

```
<seaborn.axisgrid.PairGrid at 0x7f6cbcabcf28>
```



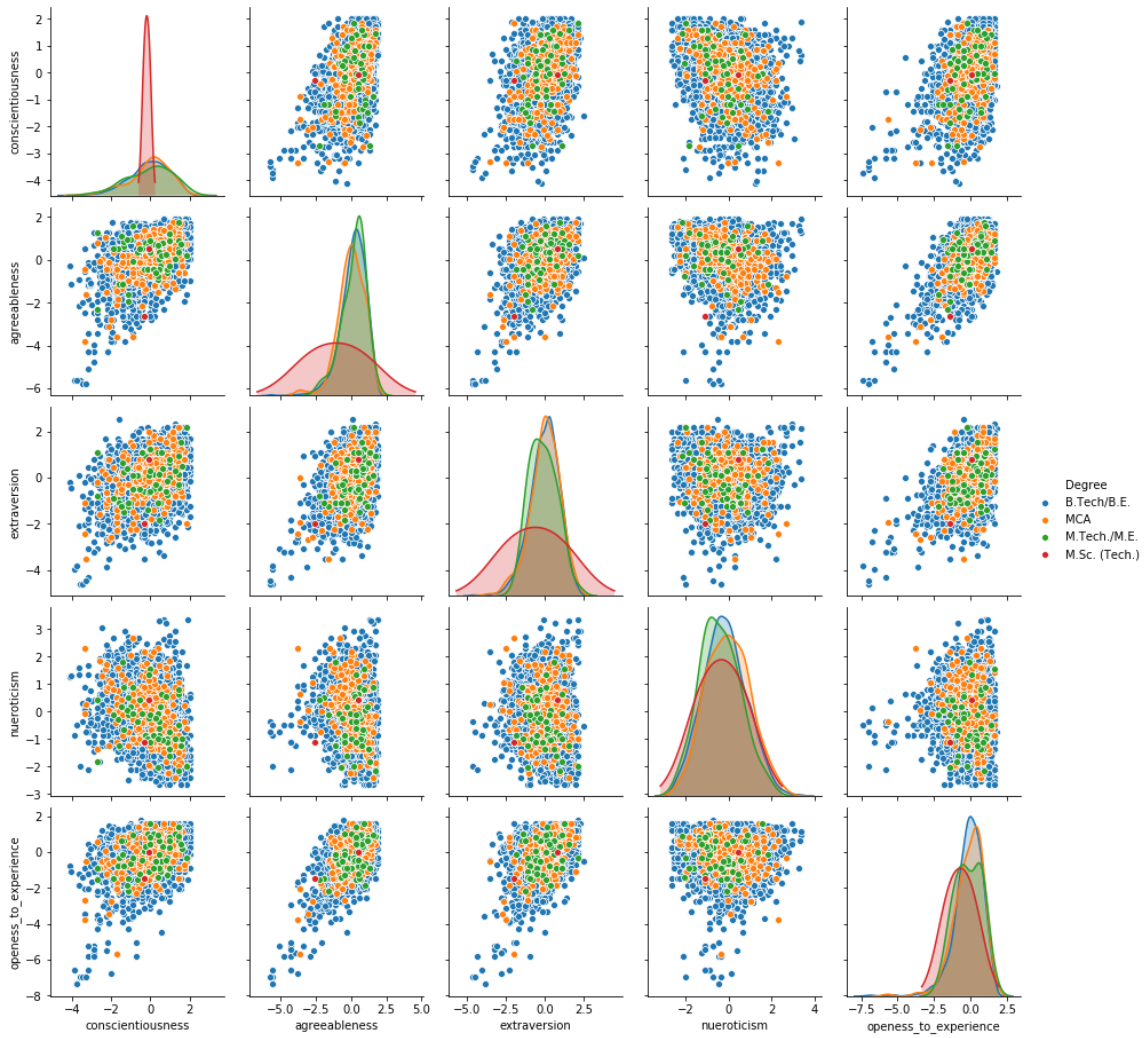
1.3.1.25 Pair plot of personality scores by Degree

In [36]:

```
personalityScores = train[['conscientiousness','agreeableness','extraversion','n
ueroticism','openess_to_experience','Degree']]
sns.pairplot(data=personalityScores,hue='Degree')
```

Out[36]:

```
<seaborn.axisgrid.PairGrid at 0x7f6cbc360c50>
```



1.4 Making necessary Assumptions and changes accordingly

- 1. As our Objective in this project is to Predict the salary given the details, We don`t need the DOJ(Date of joining),DOL(Data of leaving),JobDesignation and JobCity attributes. As they are dependent variables we can simply drop them and use salary as our only dependent variable.
- 2. We don`t need CollegeId and CollegeCityID atributes. And it is perfectly safe to drop them. Because we already have a unique ID for every candidate.
- 3. Instead of considering the specific score for each domain, We can simply consider the percentile of the domain score (Which is also a feature given).
- 4. We would encode the numerical attributes and categorical attributes and stack them to create our feature vector.

Dropping the unnecessary features

In [37]:

```
features = train.drop(['CollegeID','CollegeCityID','ComputerProgramming','Electr
onicsAndSemicon','ComputerScience','ElectricalEngg','TelecomEngg','CivilEngg','M
echanicalEngg','DOJ','DOL','Designation','JobCity','Unnamed: 0'],axis=1)
candidate_id = features['ID']
features = features.drop(['ID'],axis=1)
features.head()
```

Out[37]:

	Salary	Gender	DOB	10percentage	10board	12graduation	12percentage	12bo
0	420000	f	1990-02-19	84.3	board ofsecondary education,ap	2007	95.8	boar intermed educatior
1	500000	m	1989-10-04	85.4	cbse	2007	85.0	c
2	325000	f	1992-08-03	85.0	cbse	2010	68.2	c
3	1100000	m	1989-12-05	85.6	cbse	2007	83.6	c
4	200000	m	1991-02-27	78.0	cbse	2008	76.8	c

5 rows × 24 columns

Dividing features into dependent variables and independent variables

In [38]:

```
x = features.drop(['Salary'],axis=1)
y = features['Salary'].values
```

1.5 Dividing the data into Train,Test and Cross validation

In [39]:

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3,random_s
tate=0)
x_train, x_cv, y_train, y_cv = train_test_split(x_train, y_train, test_size=0.3,
random_state=0)
print("Shape of x_train : ",x_train.shape)
print("Shape of x_cv : ",x_cv.shape)
print("Shape of x_test : ",x_test.shape)
```

```
Shape of x_train :  (1958, 23)
Shape of x_cv :  (840, 23)
Shape of x_test :  (1200, 23)
```

1.6 Encoding the features

1.6.1 Encoding the Gender atribute

In [40]:

```
le = preprocessing.LabelEncoder()
le.fit(x_train['Gender'])
x_train_gender = le.transform(x_train['Gender']).reshape(1,-1).T
x_cv_gender = le.transform(x_cv['Gender']).reshape(1,-1).T
x_test_gender = le.transform(x_test['Gender']).reshape(1,-1).T
print(x_train_gender.shape)
print(x_cv_gender.shape)
print(x_test_gender.shape)
```

```
(1958, 1)
(840, 1)
(1200, 1)
```

1.6.2 Encoding the date of birth

Replacing the whole date of birth by only year

In [41]:

```
x_train_dob = x_train['DOB'].map(lambda x: x.year).values.reshape(1,-1).T
x_cv_dob = x_cv['DOB'].map(lambda x: x.year).values.reshape(1,-1).T
x_test_dob = x_test['DOB'].map(lambda x: x.year).values.reshape(1,-1).T
print(x_train_dob.shape)
print(x_cv_dob.shape)
print(x_test_dob.shape)
```

```
(1958, 1)
(840, 1)
(1200, 1)
```

1.6.3 Encoding the 10nth percentage

In [42]:

```
normalizer = preprocessing.Normalizer()
normalizer.fit(x_train['10percentage'].values.reshape(1,-1))
x_train_10percentage = normalizer.transform(x_train['10percentage'].values.resha
pe(1,-1)).T
x_cv_10percentage = normalizer.transform(x_cv['10percentage'].values.reshape(1,-
1)).T
x_test_10percentage = normalizer.transform(x_test['10percentage'].values.reshape
(1,-1)).T
print(x_train_10percentage.shape)
print(x_cv_10percentage.shape)
print(x_test_10percentage.shape)
```

```
(1958, 1)
(840, 1)
(1200, 1)
```

1.6.4 Encoding the 10nth board

In [43]:

```
x_train['10board'] = x_train['10board'].replace(0, 'no')
x_cv['10board'] = x_cv['10board'].replace(0, 'no')
x_test['10board'] = x_test['10board'].replace(0, 'no')
```

In [44]:

```
vectorizer = CountVectorizer()
vectorizer.fit(x_train['10board'])
x_train_10board = vectorizer.transform(x_train['10board'])
x_cv_10board = vectorizer.transform(x_cv['10board'])
x_test_10board = vectorizer.transform(x_test['10board'])
print(x_train_10board.shape)
print(x_cv_10board.shape)
print(x_test_10board.shape)
```

```
(1958, 137)
(840, 137)
(1200, 137)
```

1.6.5 Encoding the 12th percentage

In [45]:

```
normalizer = preprocessing.Normalizer()
normalizer.fit(x_train['12percentage'].values.reshape(1,-1))
x_train_12percentage = normalizer.transform(x_train['12percentage'].values.resha
pe(1,-1)).T
x_cv_12percentage = normalizer.transform(x_cv['12percentage'].values.reshape(1,-
1)).T
x_test_12percentage = normalizer.transform(x_test['12percentage'].values.reshape
(1,-1)).T
print(x_train_12percentage.shape)
print(x_cv_12percentage.shape)
print(x_test_12percentage.shape)
```

```
(1958, 1)
(840, 1)
(1200, 1)
```

1.6.6 Encoding the 12th board

In [46]:

```
x_train['12board'] = x_train['12board'].replace(0, 'no')
x_cv['12board'] = x_cv['12board'].replace(0, 'no')
x_test['12board'] = x_test['12board'].replace(0, 'no')
```

In [47]:

```
vectorizer = CountVectorizer()
vectorizer.fit(x_train['12board'])
x_train_12board = vectorizer.transform(x_train['12board'])
x_cv_12board = vectorizer.transform(x_cv['12board'])
x_test_12board = vectorizer.transform(x_test['12board'])
print(x_train_12board.shape)
print(x_cv_12board.shape)
print(x_test_12board.shape)
```

```
(1958, 183)
(840, 183)
(1200, 183)
```

1.6.7 Encoding the Degree

In [48]:

```
vectorizer = CountVectorizer()
vectorizer.fit(x_train['Degree'])
x_train_degree = vectorizer.transform(x_train['Degree'])
x_cv_degree = vectorizer.transform(x_cv['Degree'])
x_test_degree = vectorizer.transform(x_test['Degree'])
print(x_train_degree.shape)
print(x_cv_degree.shape)
print(x_test_degree.shape)
```

```
(1958, 3)
(840, 3)
(1200, 3)
```

1.6.8 Encoding the Specialization

In [49]:

```
vectorizer = CountVectorizer()
vectorizer.fit(x_train['Specialization'])
x_train_specialization = vectorizer.transform(x_train['Specialization'])
x_cv_specialization = vectorizer.transform(x_cv['Specialization'])
x_test_specialization = vectorizer.transform(x_test['Specialization'])
print(x_train_specialization.shape)
print(x_cv_specialization.shape)
print(x_test_specialization.shape)
```

```
(1958, 33)
(840, 33)
(1200, 33)
```

1.6.9 Encoding the college GPA

In [50]:

```
normalizer = preprocessing.Normalizer()
normalizer.fit(x_train['collegeGPA'].values.reshape(1,-1))
x_train_collegegpa = normalizer.transform(x_train['collegeGPA'].values.reshape(1
,-1)).T
x_cv_collegegpa = normalizer.transform(x_cv['collegeGPA'].values.reshape(1,-1)).
T
x_test_collegegpa = normalizer.transform(x_test['collegeGPA'].values.reshape(1,-
1)).T
print(x_train_collegegpa.shape)
print(x_cv_collegegpa.shape)
print(x_test_collegegpa.shape)
```

```
(1958, 1)
(840, 1)
(1200, 1)
```

1.6.10 Encoding the college state

In [51]:

```
vectorizer = CountVectorizer()
vectorizer.fit(x_train['CollegeState'])
x_train_collegestate = vectorizer.transform(x_train['CollegeState'])
x_cv_collegestate = vectorizer.transform(x_cv['CollegeState'])
x_test_collegestate = vectorizer.transform(x_test['CollegeState'])
print(x_train_collegestate.shape)
print(x_cv_collegestate.shape)
print(x_test_collegestate.shape)
```

```
(1958, 32)
(840, 32)
(1200, 32)
```

1.6.11 Encoding the English score

In [52]:

```
normalizer = preprocessing.Normalizer()
normalizer.fit(x_train['English'].values.reshape(1,-1))
x_train_english = normalizer.transform(x_train['English'].values.reshape(1,-1)).
T
x_cv_english = normalizer.transform(x_cv['English'].values.reshape(1,-1)).T
x_test_english = normalizer.transform(x_test['English'].values.reshape(1,-1)).T
print(x_train_english.shape)
print(x_cv_english.shape)
print(x_test_english.shape)
```

```
(1958, 1)
(840, 1)
(1200, 1)
```

1.6.12 Encoding the Logical score

In [53]:

```
normalizer = preprocessing.Normalizer()
normalizer.fit(x_train['Logical'].values.reshape(1,-1))
x_train_logical = normalizer.transform(x_train['Logical'].values.reshape(1,-1)).
T
x_cv_logical = normalizer.transform(x_cv['Logical'].values.reshape(1,-1)).T
x_test_logical = normalizer.transform(x_test['Logical'].values.reshape(1,-1)).T
print(x_train_logical.shape)
print(x_cv_logical.shape)
print(x_test_logical.shape)
```

```
(1958, 1)
(840, 1)
(1200, 1)
```

1.6.13 Encoding the Quants score

In [54]:

```
normalizer = preprocessing.Normalizer()
normalizer.fit(x_train['Quant'].values.reshape(1,-1))
x_train_quant = normalizer.transform(x_train['Quant'].values.reshape(1,-1)).T
x_cv_quant = normalizer.transform(x_cv['Quant'].values.reshape(1,-1)).T
x_test_quant = normalizer.transform(x_test['Quant'].values.reshape(1,-1)).T
print(x_train_quant.shape)
print(x_cv_quant.shape)
print(x_test_quant.shape)
```

```
(1958, 1)
(840, 1)
(1200, 1)
```

1.6.14 Encoding the Domain_score

In [55]:

```
x_train['Domain'] = x_train['Domain'].replace(-1, 0)
x_cv['Domain'] = x_cv['Domain'].replace(-1, 0)
x_test['Domain'] = x_test['Domain'].replace(-1, 0)
```

> Leaving the personality test scores as is, because they are sampled from a distribution with mean 0 and standard deviation 1.

1.7 Creating feature sets

In [56]:

```
x_tr = hstack((x_train_gender,x_train_dob,x_train_10percentage,x_train_10board,x_train['12graduation'].values.reshape(1,-1).T,x_train_12percentage,x_train_12board,x_train['CollegeTier'].values.reshape(1,-1).T,x_train_degree,x_train_specialization,x_train_collegegpa,x_train['CollegeCityTier'].values.reshape(1,-1).T,x_train_collegestate,x_train['GraduationYear'].values.reshape(1,-1).T,x_train_english,x_train_logical,x_train_quant,x_train['Domain'].values.reshape(1,-1).T,x_train['conscientiousness'].values.reshape(1,-1).T,x_train['agreeableness'].values.reshape(1,-1).T,x_train['extraversion'].values.reshape(1,-1).T,x_train['nueroticism'].values.reshape(1,-1).T,x_train['openess_to_experience'].values.reshape(1,-1).T)).tocsr()
x_cv = hstack((x_cv_gender,x_cv_dob,x_cv_10percentage,x_cv_10board,x_cv['12graduation'].values.reshape(1,-1).T,x_cv_12percentage,x_cv_12board,x_cv['CollegeTier'].values.reshape(1,-1).T,x_cv_degree,x_cv_specialization,x_cv_collegegpa,x_cv['CollegeCityTier'].values.reshape(1,-1).T,x_cv_collegestate,x_cv['GraduationYear'].values.reshape(1,-1).T,x_cv_english,x_cv_logical,x_cv_quant,x_cv['Domain'].values.reshape(1,-1).T,x_cv['conscientiousness'].values.reshape(1,-1).T,x_cv['agreeableness'].values.reshape(1,-1).T,x_cv['extraversion'].values.reshape(1,-1).T,x_cv['nueroticism'].values.reshape(1,-1).T,x_cv['openess_to_experience'].values.reshape(1,-1).T)).tocsr()
x_ts = hstack((x_test_gender,x_test_dob,x_test_10percentage,x_test_10board,x_test['12graduation'].values.reshape(1,-1).T,x_test_12percentage,x_test_12board,x_test['CollegeTier'].values.reshape(1,-1).T,x_test_degree,x_test_specialization,x_test_collegegpa,x_test['CollegeCityTier'].values.reshape(1,-1).T,x_test_collegestate,x_test['GraduationYear'].values.reshape(1,-1).T,x_test_english,x_test_logical,x_test_quant,x_test['Domain'].values.reshape(1,-1).T,x_test['conscientiousness'].values.reshape(1,-1).T,x_test['agreeableness'].values.reshape(1,-1).T,x_test['extraversion'].values.reshape(1,-1).T,x_test['nueroticism'].values.reshape(1,-1).T,x_test['openess_to_experience'].values.reshape(1,-1).T)).tocsr()
```

1.8 Hyperparameter tuning using Gridsearch

In [57]:

```
estimator = RandomForestRegressor()
param_grid = {
            "n_estimators"      : [10,20,30],
            "max_features"      : ["auto", "sqrt", "log2"],
            "min_samples_split" : [2,4,8],
            "bootstrap": [True, False],
            }
grid = GridSearchCV(estimator, param_grid, n_jobs=-1)
grid.fit(x_tr, y_train)
```

Out[57]:

```
GridSearchCV(cv=None, error_score=nan,
             estimator=RandomForestRegressor(bootstrap=True, ccp_alp
ha=0.0,
                                             criterion='mse', max_de
pth=None,
                                             max_features='auto',
                                             max_leaf_nodes=None,
                                             max_samples=None,
                                             min_impurity_decrease=
0.0,
                                             min_impurity_split=Non
e,
                                             min_samples_leaf=1,
                                             min_samples_split=2,
                                             min_weight_fraction_lea
f=0.0,
                                             n_estimators=100, n_job
s=None,
                                             oob_score=False, random
_state=None,
                                             verbose=0, warm_start=F
alse),
             iid='deprecated', n_jobs=-1,
             param_grid={'bootstrap': [True, False],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'min_samples_split': [2, 4, 8],
                         'n_estimators': [10, 20, 30]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score
=False,
             scoring=None, verbose=0)
```

In [58]:

```
grid.best_params_
```

Out[58]:

```
{'bootstrap': False,
 'max_features': 'sqrt',
 'min_samples_split': 8,
 'n_estimators': 30}
```

1.9 Training Random forest regressor with best hyper parameters found

In [75]:

```
rf = RandomForestRegressor(n_estimators=30,min_samples_split=4,max_features='sqr
t',bootstrap=False,criterion='mae')
rf.fit(x_tr,y_train)
```

Out[75]:

```
RandomForestRegressor(bootstrap=False, ccp_alpha=0.0, criterion='ma
e',
                      max_depth=None, max_features='sqrt', max_leaf_
nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=4, min_weight_fraction_leaf=
0.0,
                      n_estimators=30, n_jobs=None, oob_score=False,
                      random_state=None, verbose=0, warm_start=Fals
e)
```

In [76]:

```
print(rf.score(x_cv,y_cv))
print(rf.score(x_ts,y_test))
```

```
-0.6598589824536849
-0.040515210749436426
```

1.10 Applying PCA only on scores and marks/grades

In [96]:

```
x_tr_1 = np.hstack((x_train_gender,x_train_10percentage,x_train_12percentage,x_t
rain_collegegpa,x_train_english,x_train_logical,x_train_quant,x_train['Domain'].
values.reshape(1,-1).T,x_train['conscientiousness'].values.reshape(1,-1).T,x_tra
in['agreeableness'].values.reshape(1,-1).T,x_train['extraversion'].values.reshap
e(1,-1).T,x_train['nueroticism'].values.reshape(1,-1).T,x_train['openess_to_expe
rience'].values.reshape(1,-1).T))
x_ts_1 = np.hstack((x_test_gender,x_test_10percentage,x_test_12percentage,x_test
_collegegpa,x_test_english,x_test_logical,x_test_quant,x_test['Domain'].values.r
eshape(1,-1).T,x_test['conscientiousness'].values.reshape(1,-1).T,x_test['agreea
bleness'].values.reshape(1,-1).T,x_test['extraversion'].values.reshape(1,-1).T,x
_test['nueroticism'].values.reshape(1,-1).T,x_test['openess_to_experience'].valu
es.reshape(1,-1).T))
```

In [100]:

```
pca = PCA(n_components=5)
pca.fit(x_tr_1)
print(pca.explained_variance_ratio_)
print(pca.explained_variance_ratio_.sum())
```

```
[0.45949492 0.20532811 0.10938951 0.101519   0.07246479]
0.9481963339323168
```

In [101]:

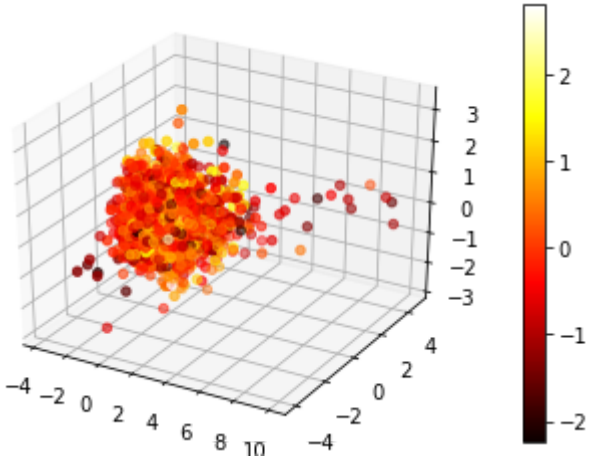
```
reduced = pca.transform(x_tr_1)
```

In [102]:

```
fig = plt.figure()
plt.figure(figsize=(50,50))
ax = fig.add_subplot(111, projection='3d')

x = reduced[:,0]
y = reduced[:,1]
z = reduced[:,2]
c = reduced[:,3]

img = ax.scatter(x, y, z, c=c, cmap=plt.hot())
fig.colorbar(img)
plt.show()
```



<Figure size 3600x3600 with 0 Axes>

1.11 Applying random forest regression on Scores feature set

hyper parameter tuning

In [103]:

```
estimator = RandomForestRegressor()
param_grid = {
            "n_estimators"      : [10,20,30],
            "max_features"      : ["auto", "sqrt", "log2"],
            "min_samples_split" : [2,4,8],
            "bootstrap": [True, False],
            }
grid = GridSearchCV(estimator, param_grid, n_jobs=-1)
grid.fit(x_tr_1, y_train)
```

Out[103]:

```
GridSearchCV(cv=None, error_score=nan,
             estimator=RandomForestRegressor(bootstrap=True, ccp_alp
ha=0.0,
                                             criterion='mse', max_de
pth=None,
                                             max_features='auto',
                                             max_leaf_nodes=None,
                                             max_samples=None,
                                             min_impurity_decrease=
0.0,
                                             min_impurity_split=Non
e,
                                             min_samples_leaf=1,
                                             min_samples_split=2,
                                             min_weight_fraction_lea
f=0.0,
                                             n_estimators=100, n_job
s=None,
                                             oob_score=False, random
_state=None,
                                             verbose=0, warm_start=F
alse),
             iid='deprecated', n_jobs=-1,
             param_grid={'bootstrap': [True, False],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'min_samples_split': [2, 4, 8],
                         'n_estimators': [10, 20, 30]},
             pre_dispatch='2*n_jobs', refit=True, return_train_score
=False,
             scoring=None, verbose=0)
```

In [104]:

```
grid.best_params_
```

Out[104]:

```
{'bootstrap': True,
 'max_features': 'log2',
 'min_samples_split': 8,
 'n_estimators': 30}
```

In [74]:

```
rf = RandomForestRegressor(n_estimators=30,min_samples_split=8,max_features='log
2',bootstrap=True,criterion='mae')
rf.fit(x_tr_1,y_train)
rf.score(x_tr_1,y_train)
```

Out[74]:

```
0.7434388149760832
```

1.12 storing pickle into a file

In [108]:

```
from sklearn.externals import joblib
joblib.dump(rf, 'rf.pkl')
```

Out[108]:

```
['rf.pkl']
```