In [25]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
mData = pd.read_csv('Data Sets/Sloan Digital Sky Survey DR14/Skyserver_SQL2_27_2018 6_51_39 PM.csv')
#printing the shape of the dataset
print('The Shape of The Data ',mData.shape)
```

```
The Shape of The Data  (10000, 18)
```

In [4]:
```
mData.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 18 columns):
objid        10000 non-null float64
ra           10000 non-null float64
dec          10000 non-null float64
u            10000 non-null float64
g            10000 non-null float64
r            10000 non-null float64
i            10000 non-null float64
z            10000 non-null float64
run          10000 non-null int64
rerun        10000 non-null int64
camcol       10000 non-null int64
field        10000 non-null int64
specobjid    10000 non-null float64
class        10000 non-null object
redshift     10000 non-null float64
plate        10000 non-null int64
mjd          10000 non-null int64
fiberid      10000 non-null int64
dtypes: float64(10), int64(7), object(1)
memory usage: 1.4+ MB
```

In [23]:
```
Y=mData['class']
X=mData.drop(columns=['class','objid','rerun'])
X.shape
```

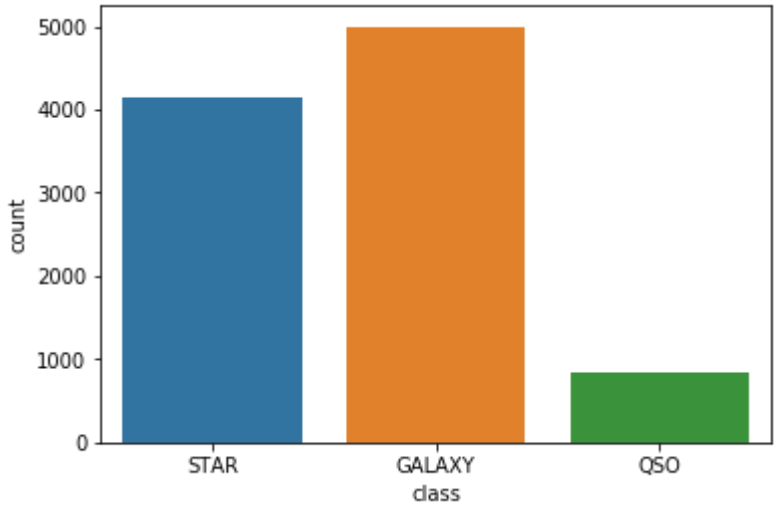
Out[23]:
```
(10000, 15)
```

Data Visualization

In [16]:
```
#Target Distribution
print(Y.value_counts())
sns.countplot(Y)
```

```
GALAXY    4998
STAR      4152
QSO        850
Name: class, dtype: int64
```

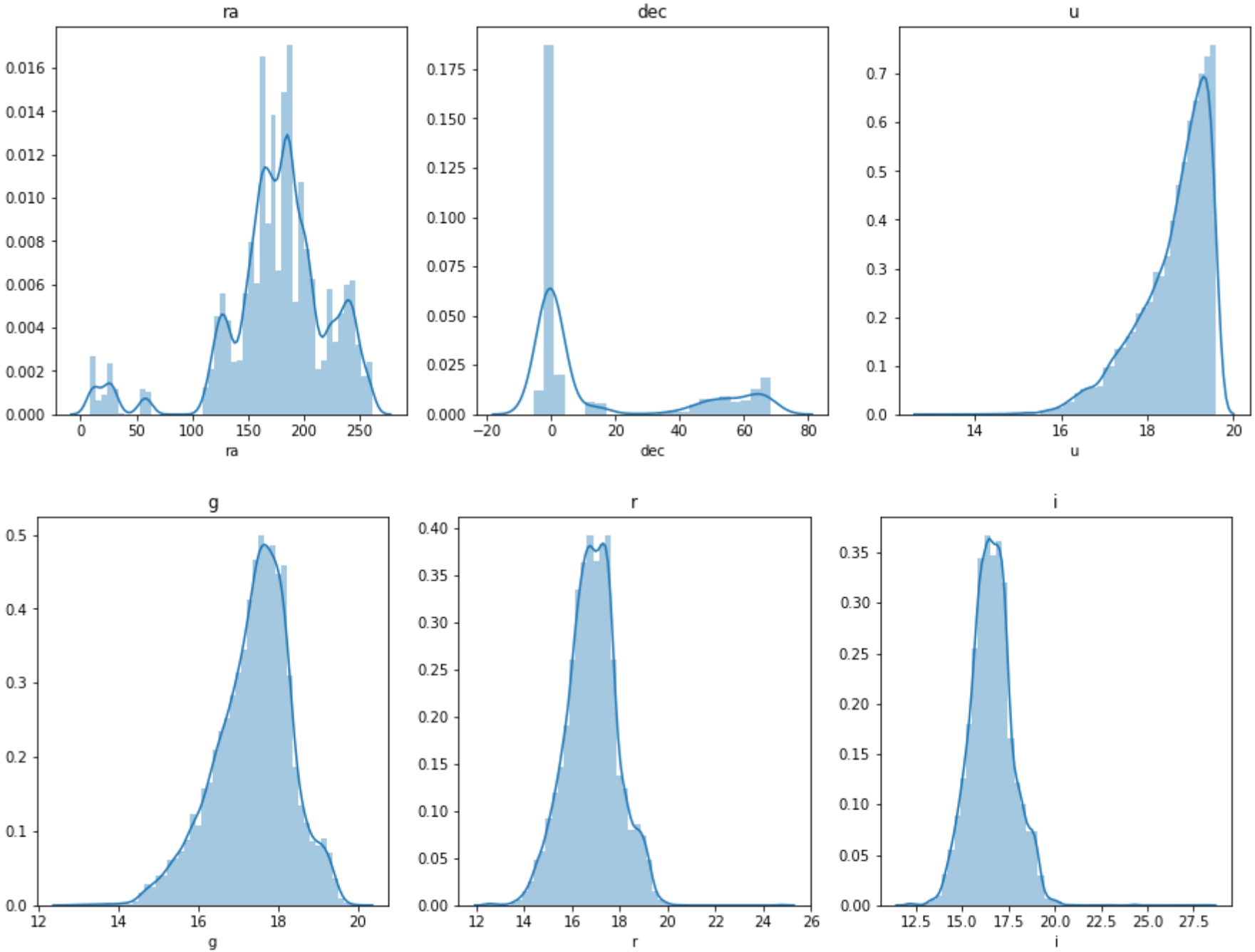
Out[16]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x110449320>
```

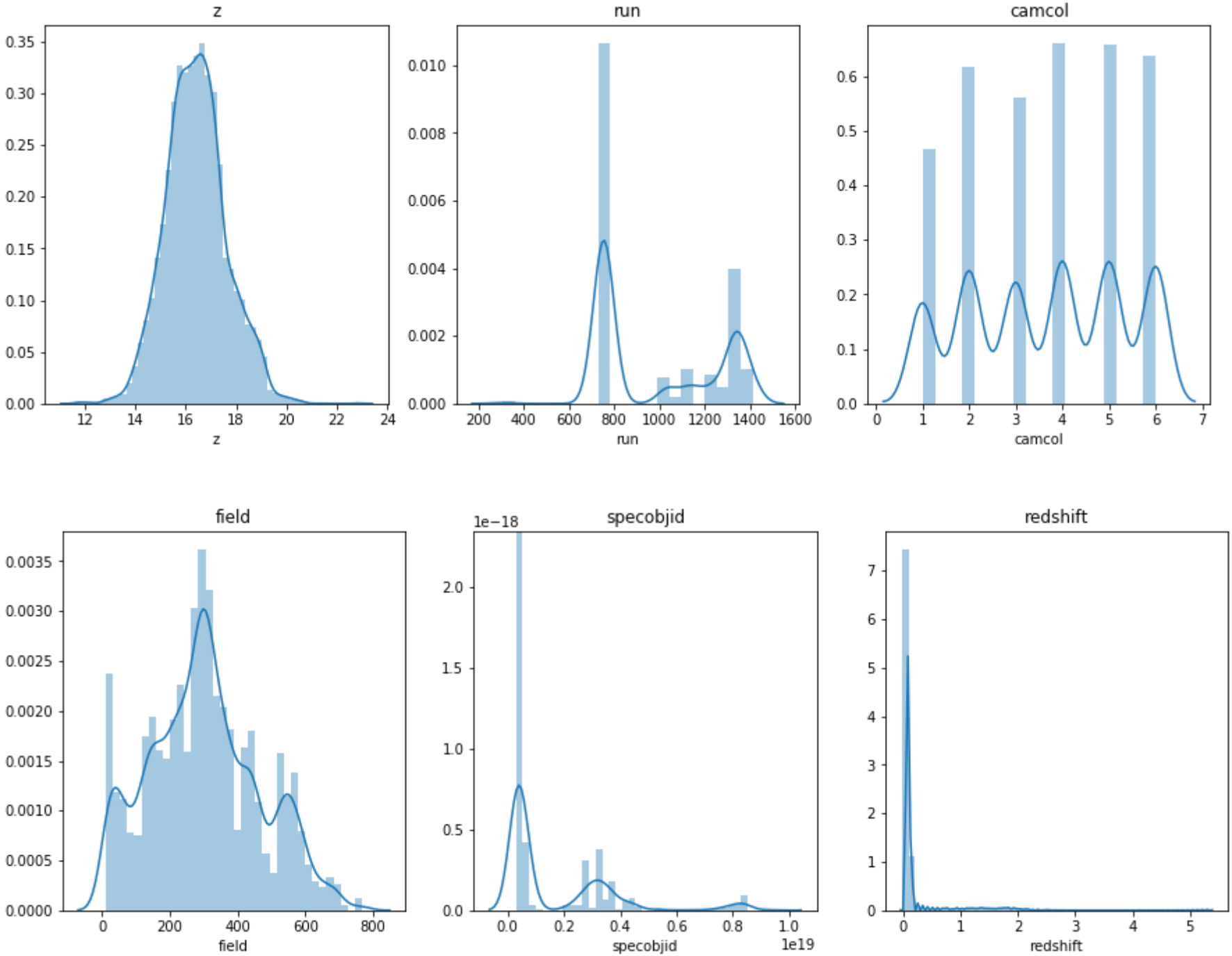


In [26]:
```
#histogram for all features
plt.figure(1,figsize=[15,5])
plt.subplot(1,3,1)
sns.distplot(X.ra)
plt.title("ra")


plt.figure(1,figsize=[15,5])
plt.subplot(1,3,2)
sns.distplot(X.dec)
plt.title("dec")

plt.figure(1,figsize=[15,5])
plt.subplot(1,3,3)
sns.distplot(X.u)
plt.title("u")

plt.figure(2,figsize=[15,5])
plt.subplot(1,3,1)
sns.distplot(X.g)
plt.title("g")


plt.figure(2,figsize=[15,5])
plt.subplot(1,3,2)
sns.distplot(X.r)
plt.title("r")


plt.figure(2,figsize=[15,5])
plt.subplot(1,3,3)
sns.distplot(X.i)
plt.title("i")

plt.figure(3,figsize=[15,5])
plt.subplot(1,3,1)
sns.distplot(X.z)
plt.title("z")


plt.figure(3,figsize=[15,5])
```

```
plt.subplot(1,3,2)
sns.distplot(X.run)
plt.title("run")


plt.figure(3,figsize=[15,5])
plt.subplot(1,3,3)
sns.distplot(X.camcol)
plt.title("camcol")

plt.figure(4,figsize=[15,5])
plt.subplot(1,3,1)
sns.distplot(X.field)
plt.title("field")

plt.figure(4,figsize=[15,5])
plt.subplot(1,3,2)
sns.distplot(X["specobjid"])
plt.title("specobjid")

plt.figure(4,figsize=[15,5])
plt.subplot(1,3,3)
sns.distplot(X.redshift)
plt.title("redshift")


plt.figure(5,figsize=[15,5])
plt.subplot(1,3,1)
sns.distplot(X.plate)
plt.title("plate")

plt.figure(5,figsize=[15,5])
plt.subplot(1,3,2)
sns.distplot(X.mjd)
plt.title("mjd")

plt.figure(5,figsize=[15,5])
plt.subplot(1,3,3)
sns.distplot(X.fiberid)
plt.title("fiberid")
```

Out[26]: Text(0.5,1,'fiberid')




z
run
camcol
field
specobjid
1e−18
1e19
redshift

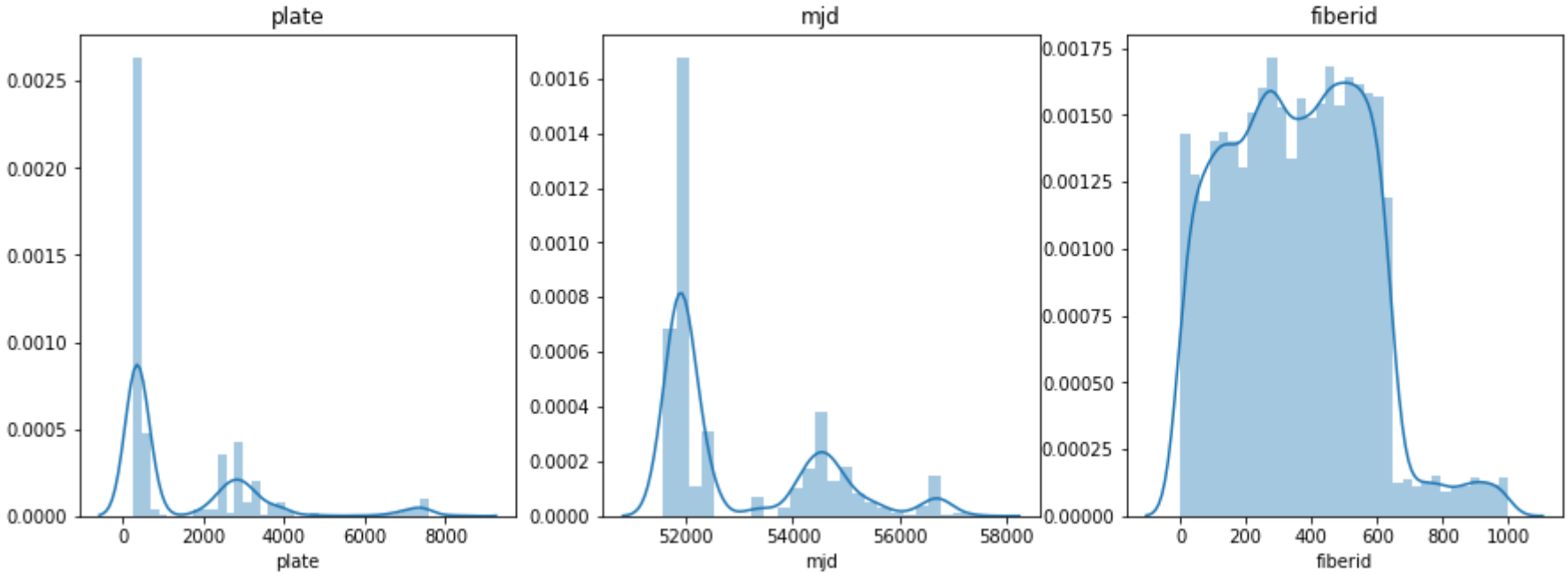


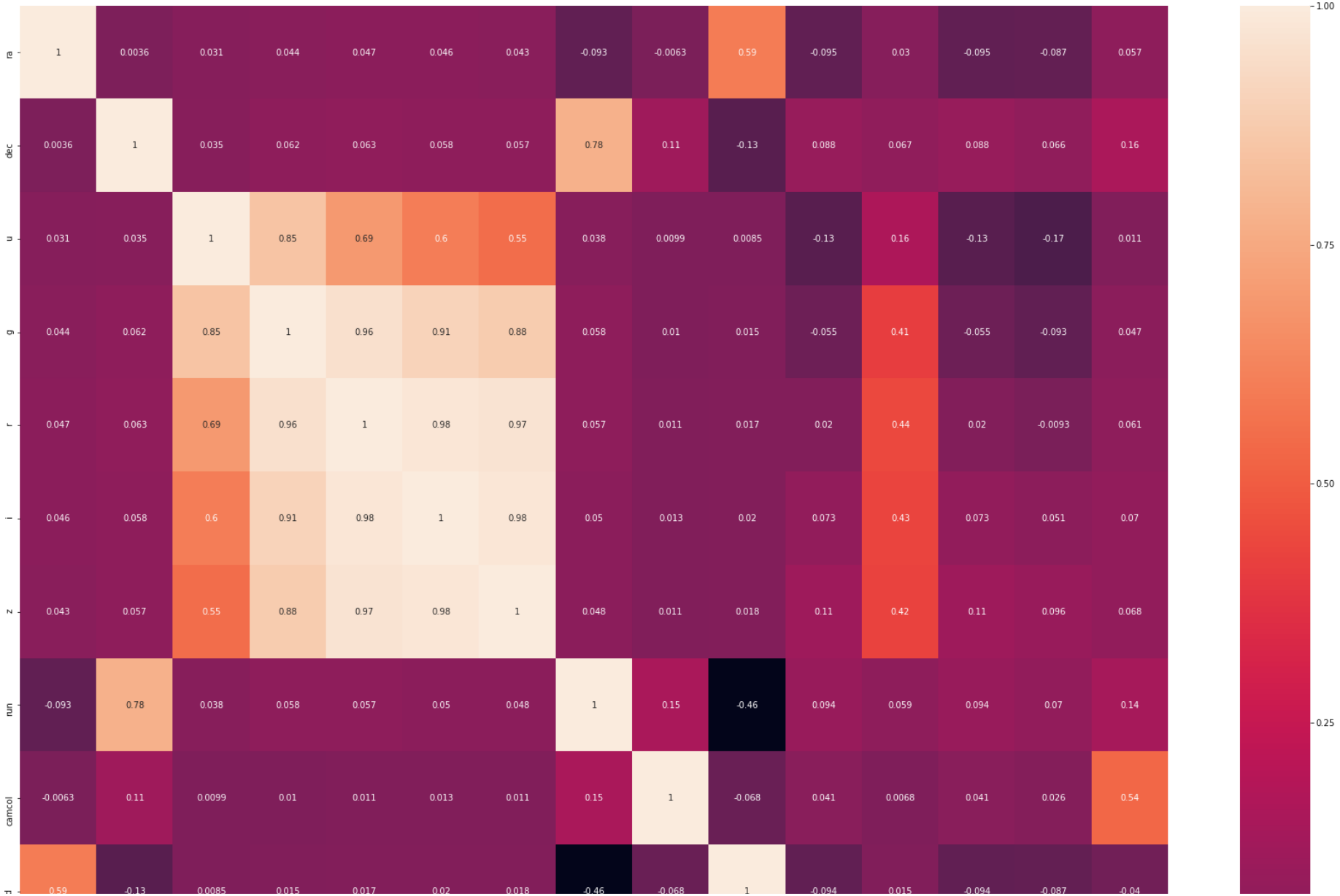
plate
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
2000
4000
6000
8000
plate
mjd
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
52000
54000
56000
58000
mjd
fiberid
0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
0
200
400
600
800
1000
fiberid

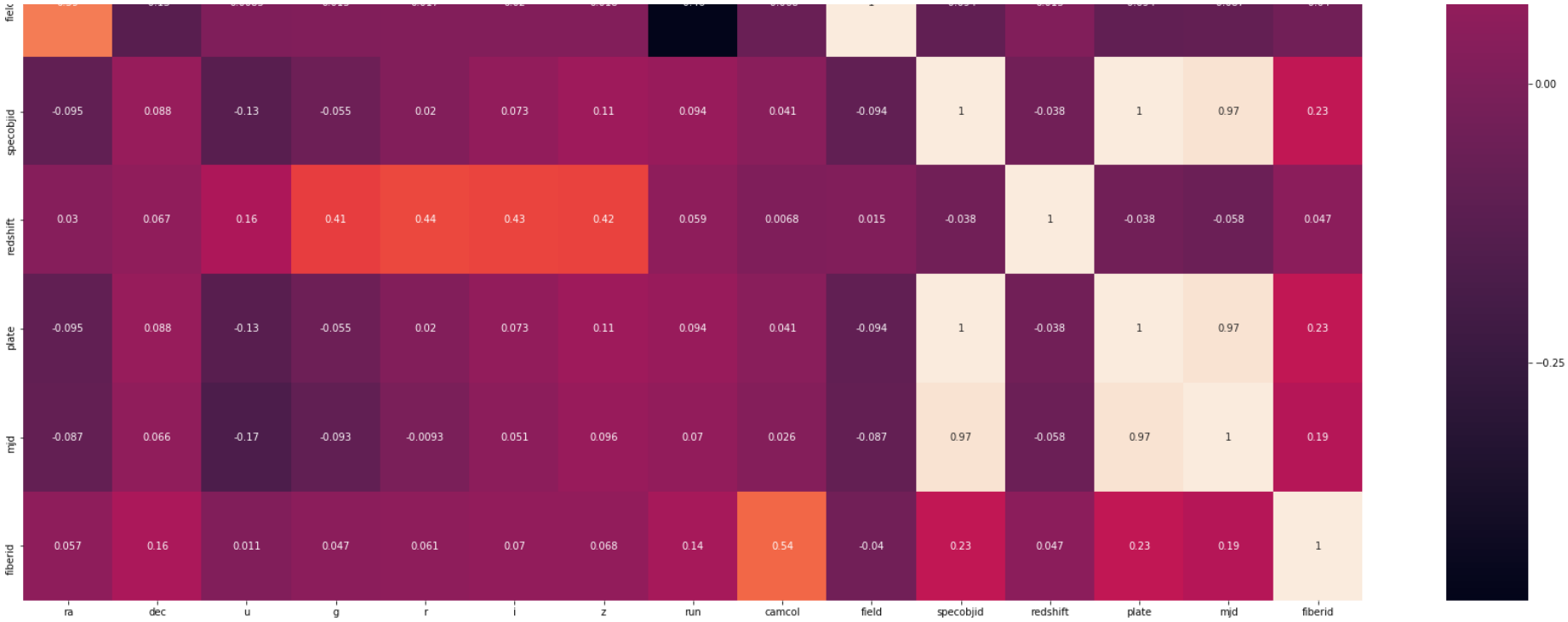
In [38]:
```
print(X.describe())
```

```
                ra           dec             u             g             r  \
count  10000.000000  10000.000000  10000.000000  10000.000000  10000.000000
mean     175.529987     14.836148     18.619355     17.371931     16.840963
std       47.783439     25.212207      0.828656      0.945457      1.067764
min        8.235100     -5.382632     12.988970     12.799550     12.431600
25%      157.370946     -0.539035     18.178035     16.815100     16.173333
50%      180.394514      0.404166     18.853095     17.495135     16.858770
75%      201.547279     35.649397     19.259232     18.010145     17.512675
max      260.884382     68.542265     19.599900     19.918970     24.802040

                i             z           run        camcol         field  \
count  10000.000000  10000.000000  10000.000000  10000.000000  10000.000000
mean      16.583579     16.422833    981.034800      3.648700    302.380100
std        1.141805      1.203188    273.305024      1.666183    162.577763
min       11.947210     11.610410    308.000000      1.000000     11.000000
25%       15.853705     15.618285    752.000000      2.000000    184.000000
50%       16.554985     16.389945    756.000000      4.000000    299.000000
75%       17.258550     17.141447   1331.000000      5.000000    414.000000
max       28.179630     22.833060   1412.000000      6.000000    768.000000

           specobjid      redshift         plate           mjd       fiberid
count   1.000000e+04  10000.000000  10000.000000  10000.000000  10000.000000
mean    1.645022e+18      0.143726   1460.986400  52943.533300    353.069400
std     2.013998e+18      0.388774   1788.778371   1511.150651    206.298149
min     2.995780e+17     -0.004136    266.000000  51578.000000      1.000000
25%     3.389248e+17      0.000081    301.000000  51900.000000    186.750000
50%     4.966580e+17      0.042591    441.000000  51997.000000    351.000000
75%     2.881300e+18      0.092579   2559.000000  54468.000000    510.000000
max     9.468830e+18      5.353854   8410.000000  57481.000000   1000.000000
```

In [27]:
```
#Heat map for to understand the correlation between the features
plt.figure(figsize=[30,30])
sns.heatmap(X.corr(), annot = True)
```

Out[27]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1d8fc4a8>

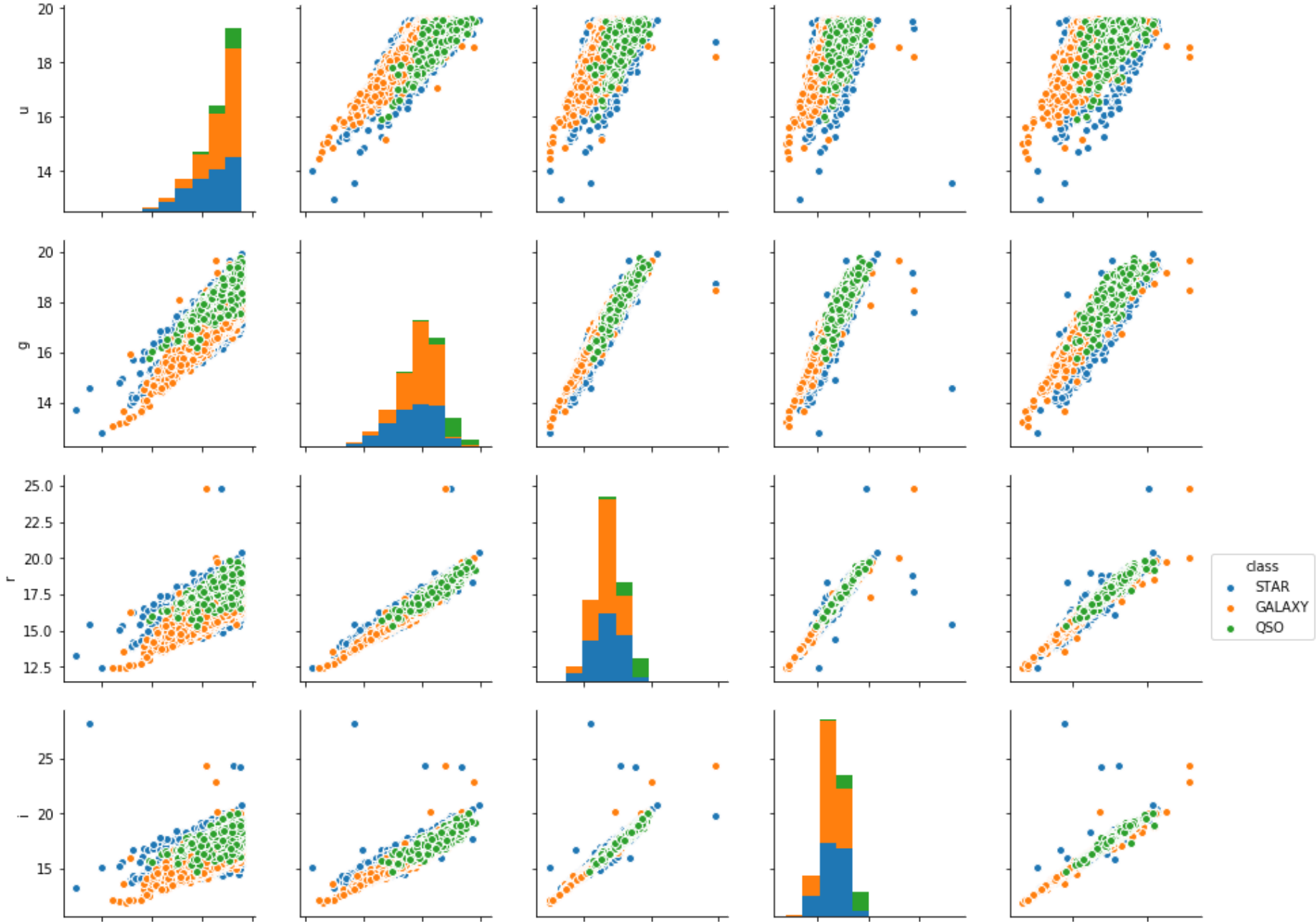


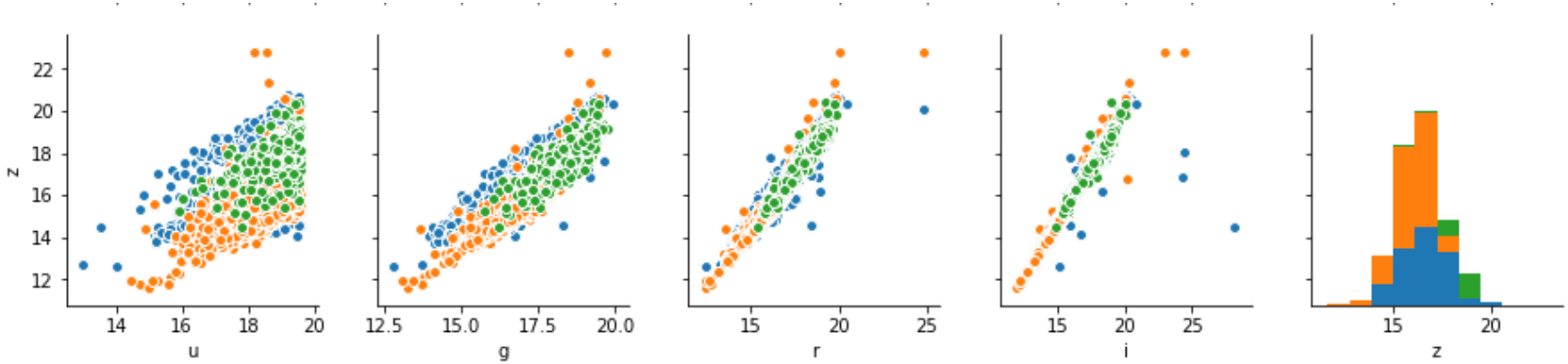


Drawing the reation between the highly corelated features

In [28]:
```
sns.pairplot(mData[["u", "g", "r","i","z","class"]],hue='class')
```

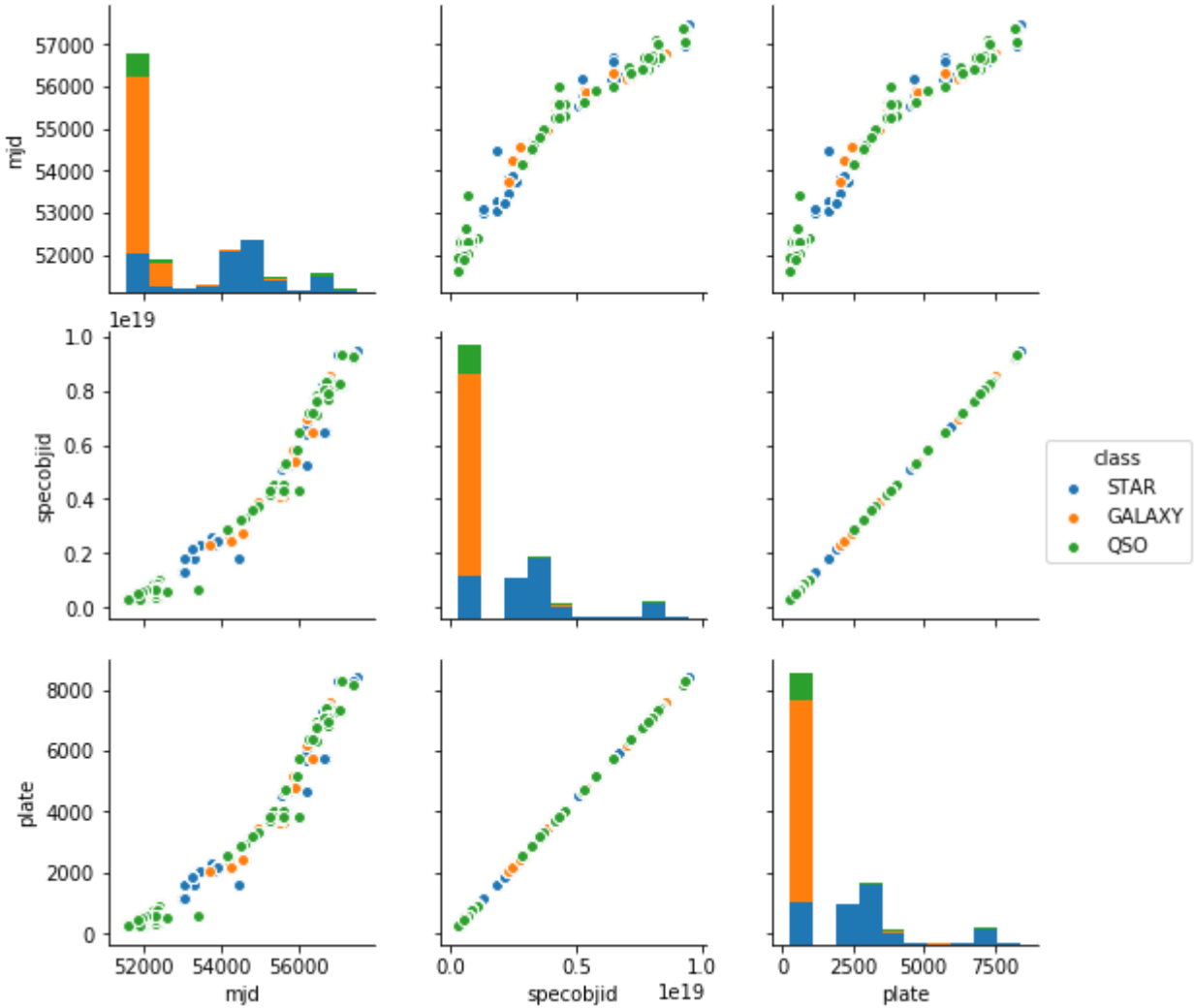
Out[28]: <seaborn.axisgrid.PairGrid at 0x1a1e079fd0>



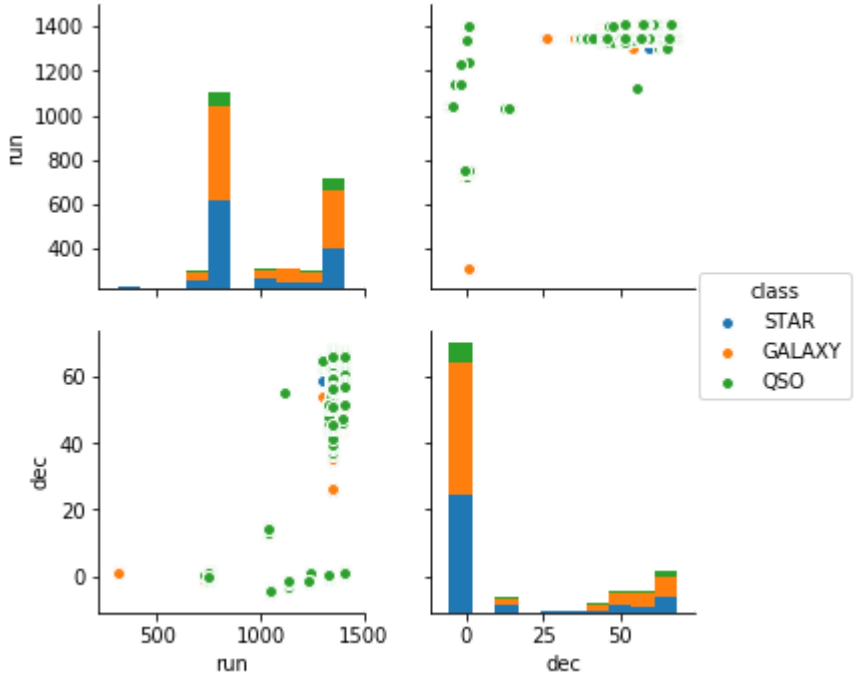

z
22
20
18
16
14
12
14
16
18
20
u
12.5
15.0
17.5
20.0
g
15
20
25
r
15
20
25
i
15
20
z

In [33]:
```
sns.pairplot(mData[["mjd", "specobjid",'plate',"class"]],hue='class')
```

Out[33]: <seaborn.axisgrid.PairGrid at 0x1a20d1fe10>

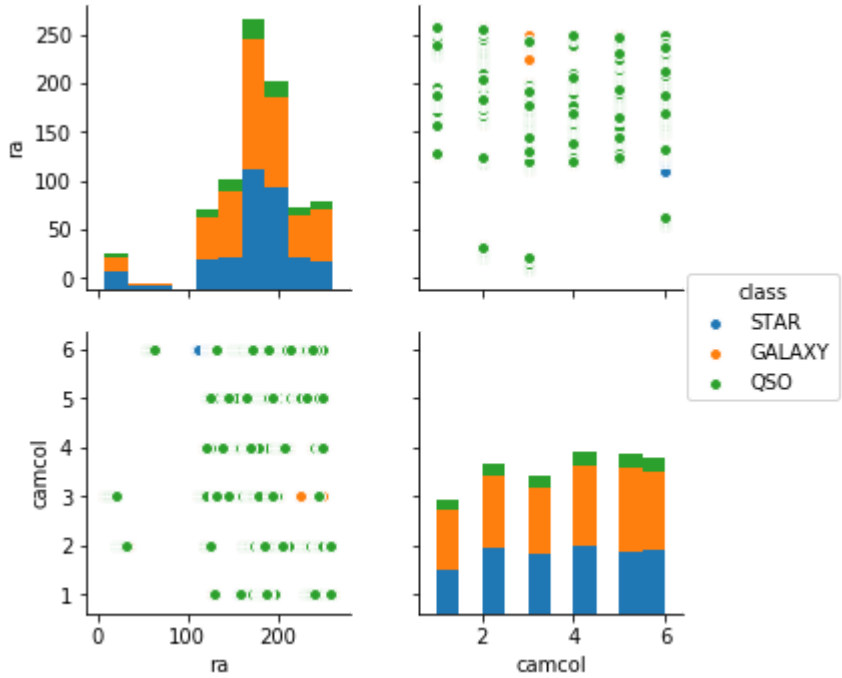


In [40]:
```
n=sns.pairplot(mData[["run", "dec","class"]],hue='class')
```



In [36]:
```
sns.pairplot(mData[["ra", "camcol","class"]],hue='class')
```

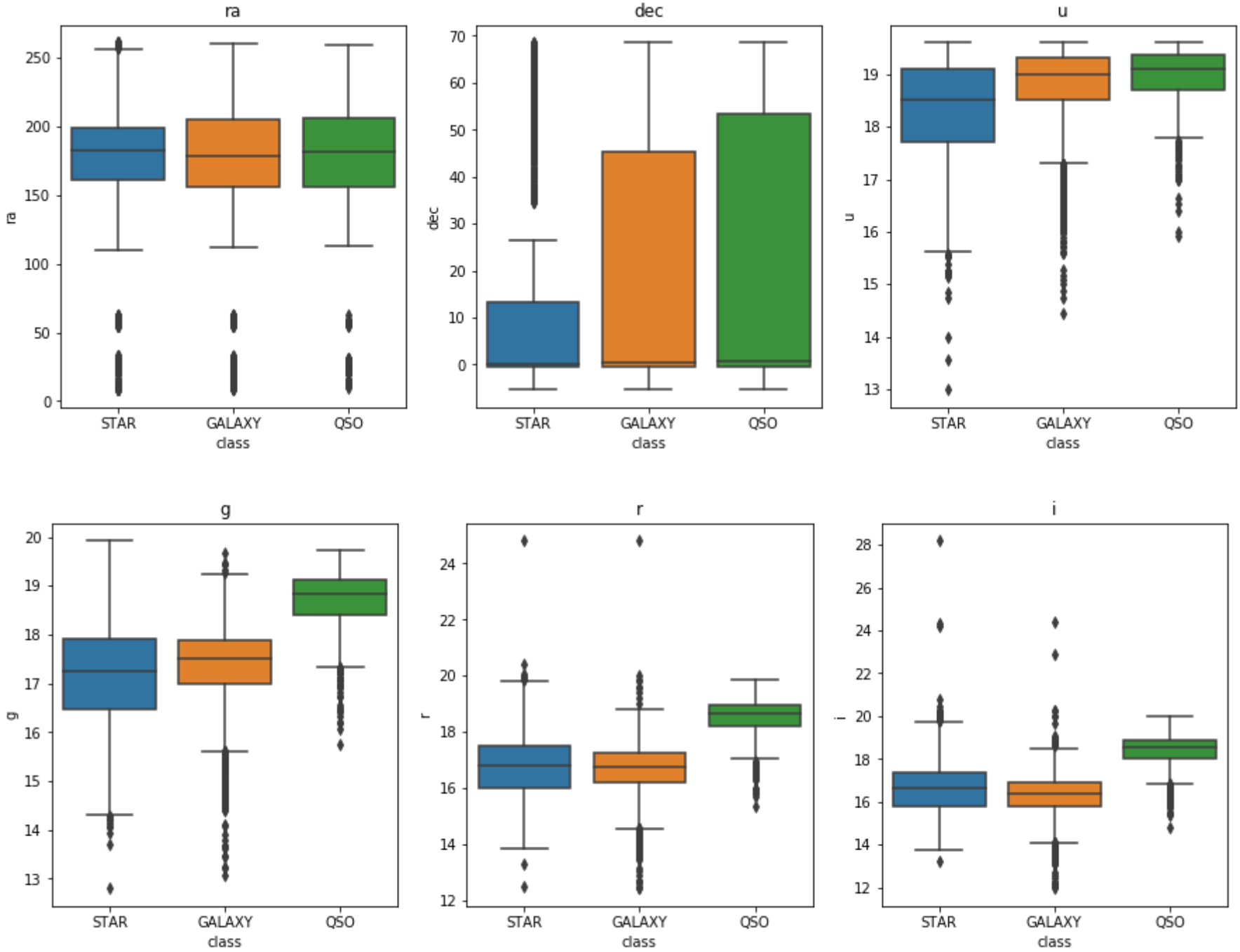
Out[36]:
```
<seaborn.axisgrid.PairGrid at 0x1a20af34e0>
```

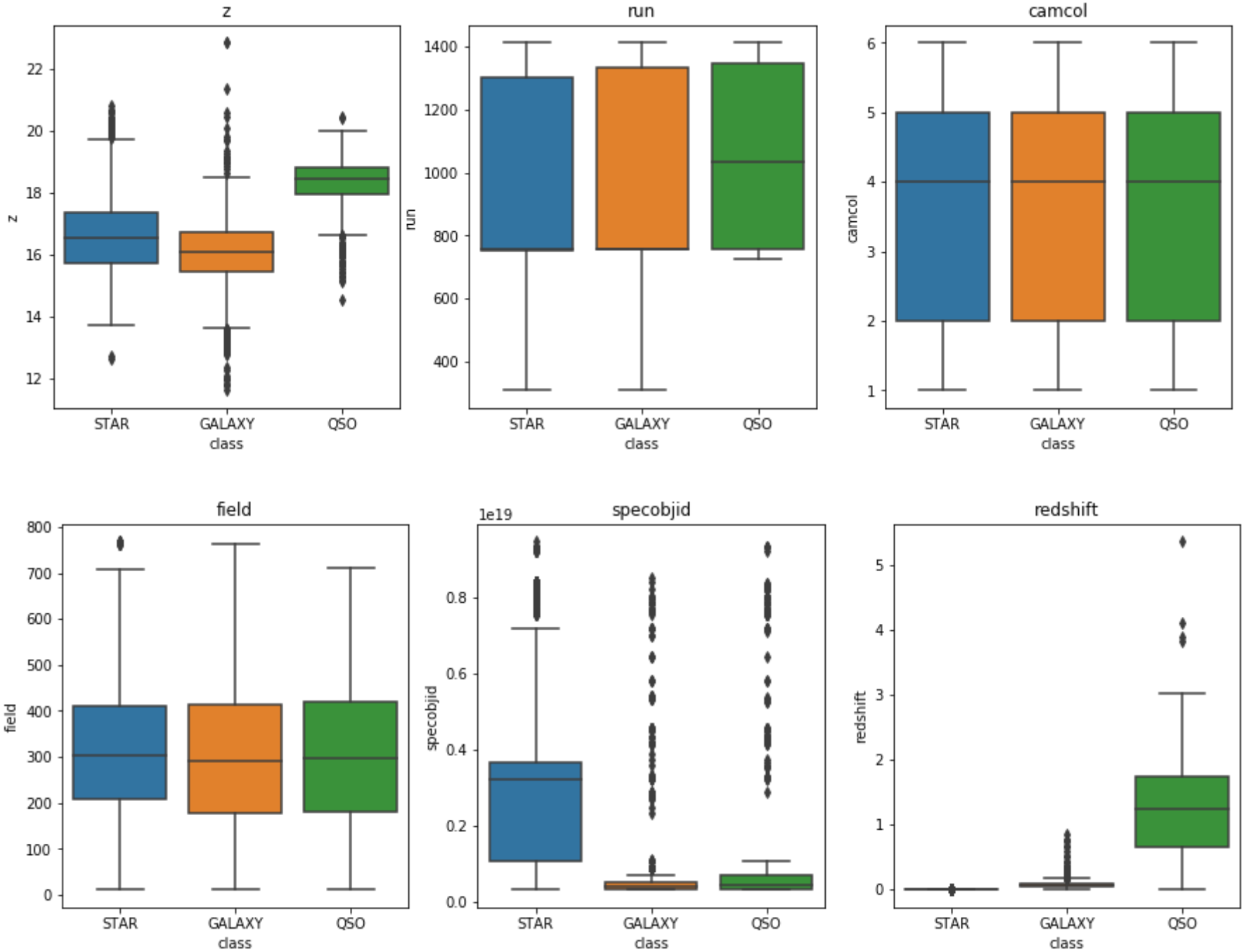


In [32]:
```
#histogram for numeric attributes
plt.figure(1,figsize=[15,5])
plt.subplot(1,3,1)
sns.boxplot(x = Y, y = X['ra'])
plt.title("ra")


plt.figure(1,figsize=[15,5])
plt.subplot(1,3,2)
sns.boxplot(x = Y, y = X['dec'])
plt.title("dec")

plt.figure(1,figsize=[15,5])
plt.subplot(1,3,3)
sns.boxplot(x = Y, y = X['u'])
plt.title("u")

plt.figure(2,figsize=[15,5])
plt.subplot(1,3,1)
sns.boxplot(x = Y, y = X['g'])
plt.title("g")


plt.figure(2,figsize=[15,5])
plt.subplot(1,3,2)
sns.boxplot(x = Y, y = X['r'])
plt.title("r")


plt.figure(2,figsize=[15,5])
plt.subplot(1,3,3)
sns.boxplot(x = Y, y = X['i'])
plt.title("i")

plt.figure(3,figsize=[15,5])
plt.subplot(1,3,1)
sns.boxplot(x = Y, y = X['z'])
plt.title("z")


plt.figure(3,figsize=[15,5])
```

```
plt.subplot(1,3,2)
sns.boxplot(x = Y, y = X['run'])
plt.title("run")


plt.figure(3,figsize=[15,5])
plt.subplot(1,3,3)
sns.boxplot(x = Y, y = X['camcol'])
plt.title("camcol")

plt.figure(4,figsize=[15,5])
plt.subplot(1,3,1)
sns.boxplot(x = Y, y = X['field'])
plt.title("field")

plt.figure(4,figsize=[15,5])
plt.subplot(1,3,2)
sns.boxplot(x = Y, y = X['specobjid'])
plt.title("specobjid")

plt.figure(4,figsize=[15,5])
plt.subplot(1,3,3)
sns.boxplot(x = Y, y = X['redshift'])
plt.title("redshift")

plt.figure(5,figsize=[15,5])
plt.subplot(1,3,1)
sns.boxplot(x = Y, y = X['plate'])
plt.title("plate")

plt.figure(5,figsize=[15,5])
plt.subplot(1,3,2)
sns.boxplot(x = Y, y = X['mjd'])
plt.title("mjd")

plt.figure(5,figsize=[15,5])
plt.subplot(1,3,3)
sns.boxplot(x = Y, y = X['fiberid'])
plt.title("fiberid")
```

Out[32]:
```
Text(0.5,1,'fiberid')
```


ra
dec
u
g
r
i
STAR
GALAXY
QSO
class


z
run
camcol
field
specobjid
1e19
redshift
STAR
GALAXY
QSO
class

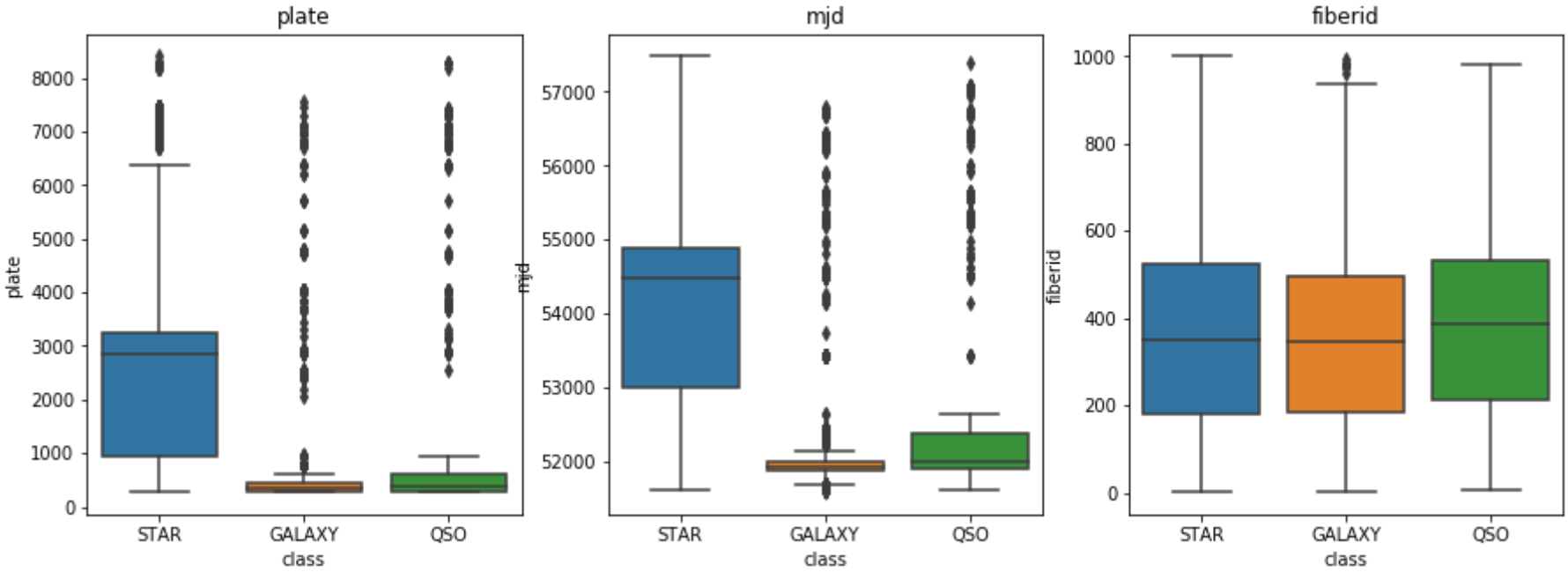


plate
mjd
fiberid
plate
mjd
fiberid
class
class
class
STAR
GALAXY
QSO
0
1000
2000
3000
4000
5000
6000
7000
8000
52000
53000
54000
55000
56000
57000
0
200
400
600
800
1000

In [ ]: