Indian School Education Statistics (Part-1)

Data Analysis of Dropout Ratio in India during year 2012-2015.

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
dropout_df = pd.read_csv('dropout-ratio-2012-2015.csv')
```

In [3]:
```
dropout_df.head()
```

Out[3]:

	State_UT	year	Primary_Boys	Primary_Girls	Primary_Total	Upper Primary_Boys	Upper Primary_Girls	I
0	A & N Islands	2012-13	0.83	0.51	0.68	Uppe_r_Primary	1.09	
1	A & N Islands	2013-14	1.35	1.06	1.21	NR	1.54	
2	A & N Islands	2014-15	0.47	0.55	0.51	1.44	1.95	
3	Andhra Pradesh	2012-13	3.3	3.05	3.18	3.21	3.51	
4	Andhra Pradesh	2013-14	4.31	4.39	4.35	3.46	4.12	

In [4]:
```
dropout_df.tail()
```

Out[4]:

	State_UT	year	Primary_Boys	Primary_Girls	Primary_Total	Upper Primary_Boys	Upper Primary_Girls
105	West Bengal	2013-14	3.44	2.37	2.91	5.63	3.1
106	West Bengal	2014-15	2.13	0.79	1.47	5.84	2.88
107	All India	2012-13	4.68	4.66	4.67	2.3	4.01
108	All India	2013-14	4.53	4.14	4.34	3.09	4.49
109	All India	2014-15	4.36	3.88	4.13	3.49	4.6

In [5]:
```
dropout_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110 entries, 0 to 109
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   State_UT             110 non-null    object
 1   year                 110 non-null    object
 2   Primary_Boys         110 non-null    object
 3   Primary_Girls        110 non-null    object
 4   Primary_Total        110 non-null    object
 5   Upper Primary_Boys   110 non-null    object
 6   Upper Primary_Girls  110 non-null    object
 7   Upper Primary_Total  110 non-null    object
 8   Secondary _Boys      110 non-null    object
 9   Secondary _Girls     110 non-null    object
 10  Secondary _Total     110 non-null    object
 11  HrSecondary_Boys     110 non-null    object
 12  HrSecondary_Girls    110 non-null    object
 13  HrSecondary_Total    110 non-null    object
dtypes: object(14)
memory usage: 12.2+ KB
```

In [6]:
```
dropout_df.isna().sum()
```

Out[6]:
```
State_UT              0
year                  0
Primary_Boys          0
Primary_Girls         0
Primary_Total         0
Upper Primary_Boys    0
Upper Primary_Girls   0
Upper Primary_Total   0
Secondary _Boys       0
Secondary _Girls      0
Secondary _Total      0
HrSecondary_Boys      0
HrSecondary_Girls     0
HrSecondary_Total     0
dtype: int64
```

In [7]:
```
dropout_df['State_UT'].unique()
```

Out[7]:
```
array(['A & N Islands', 'Andhra Pradesh', 'Arunachal  Pradesh',
       'Arunachal Pradesh', 'Assam', 'Bihar', 'Chandigarh',
       'Chhattisgarh', 'Dadra & Nagar Haveli', 'Daman & Diu', 'Delhi',
       'Goa', 'Gujarat', 'Haryana', 'Himachal Pradesh', 'Jammu & Kashmir',
       'Jharkhand', 'Karnataka', 'Kerala', 'Lakshadweep',
       'Madhya  Pradesh', 'Madhya Pradesh', 'Maharashtra', 'Manipur',
       'Meghalaya', 'Mizoram', 'Nagaland', 'Odisha', 'Puducherry',
       'Punjab', 'Rajasthan', 'Sikkim', 'Tamil  Nadu', 'Tamil Nadu',
       'Telangana', 'Tripura', 'Uttar Pradesh', 'Uttarakhand',
       'West Bengal', 'All India'], dtype=object)
```

Data Preprocessing

In [8]:
```
dropout_df.replace('NR',0,inplace=True)
```

In [9]:
```
dropout_df.replace('Uppe_r_Primary',0,inplace=True)
```

In [10]:
```
# Converting datatype of column (Primary Boys and Girls) from object to float.

dropout_df['Primary_Boys'] = dropout_df['Primary_Boys'].apply(lambda x: float(
dropout_df['Primary_Girls'] = dropout_df['Primary_Girls'].apply(lambda x: floa
```

In [11]:
```
# Converting datatype of column (Upper Primary Boys and Girls) from object to

dropout_df['Upper Primary_Boys'] = dropout_df['Upper Primary_Boys'].apply(lamb
dropout_df['Upper Primary_Girls'] = dropout_df['Upper Primary_Girls'].apply(la
```

In [12]:
```
# Converting datatype of column (Secondary Boys and Girls) from object to floa

dropout_df['Secondary _Boys'] = dropout_df['Secondary _Boys'].apply(lambda x:
dropout_df['Secondary _Girls'] = dropout_df['Secondary _Girls'].apply(lambda x
```

In [13]:
```
# Converting datatype of column (HrSecondary Boys and Girls) from object to fl

dropout_df['HrSecondary_Boys'] = dropout_df['HrSecondary_Boys'].apply(lambda x
dropout_df['HrSecondary_Girls'] = dropout_df['HrSecondary_Girls'].apply(lambda
```

In [14]:
```
# Converting datatype of column ( Total of Primary, Upper Primary, Secondary a
# and Girls) from object to float.

# Primary Total
dropout_df['Primary_Total'] = dropout_df['Primary_Total'].apply(lambda x: floa

# Upper Primary Total
dropout_df['Upper Primary_Total'] = dropout_df['Upper Primary_Total'].apply(la

# Secondary Total
dropout_df['Secondary _Total'] = dropout_df['Secondary _Total'].apply(lambda x

#HrSecondary Total
dropout_df['HrSecondary_Total'] = dropout_df['HrSecondary_Total'].apply(lambda
```

In [15]:
```
#Check info() of dataframe
dropout_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110 entries, 0 to 109
Data columns (total 14 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   State_UT             110 non-null    object 
 1   year                 110 non-null    object 
 2   Primary_Boys         110 non-null    float64
 3   Primary_Girls        110 non-null    float64
 4   Primary_Total        110 non-null    float64
 5   Upper Primary_Boys   110 non-null    float64
 6   Upper Primary_Girls  110 non-null    float64
 7   Upper Primary_Total  110 non-null    float64
 8   Secondary _Boys      110 non-null    float64
 9   Secondary _Girls     110 non-null    float64
 10  Secondary _Total     110 non-null    float64
 11  HrSecondary_Boys     110 non-null    float64
 12  HrSecondary_Girls    110 non-null    float64
 13  HrSecondary_Total    110 non-null    float64
dtypes: float64(12), object(2)
memory usage: 12.2+ KB
```

In [16]:
```
dropout_df.replace('Arunachal  Pradesh','Arunachal Pradesh',inplace=True)

dropout_df.replace('Madhya  Pradesh','Madhya Pradesh',inplace=True)

dropout_df.replace('Tamil  Nadu','Tamil Nadu',inplace=True)
```

In [17]:
```
dropout_df.head()
```

Out[17]:

	State_UT	year	Primary_Boys	Primary_Girls	Primary_Total	Upper Primary_Boys	Upper Primary_Girls	P
0	A & N Islands	2012-13	0.83	0.51	0.68	0.00	1.09	
1	A & N Islands	2013-14	1.35	1.06	1.21	0.00	1.54	
2	A & N Islands	2014-15	0.47	0.55	0.51	1.44	1.95	
3	Andhra Pradesh	2012-13	3.30	3.05	3.18	3.21	3.51	
4	Andhra Pradesh	2013-14	4.31	4.39	4.35	3.46	4.12	

In [18]:
```
dropout_df.describe()
```

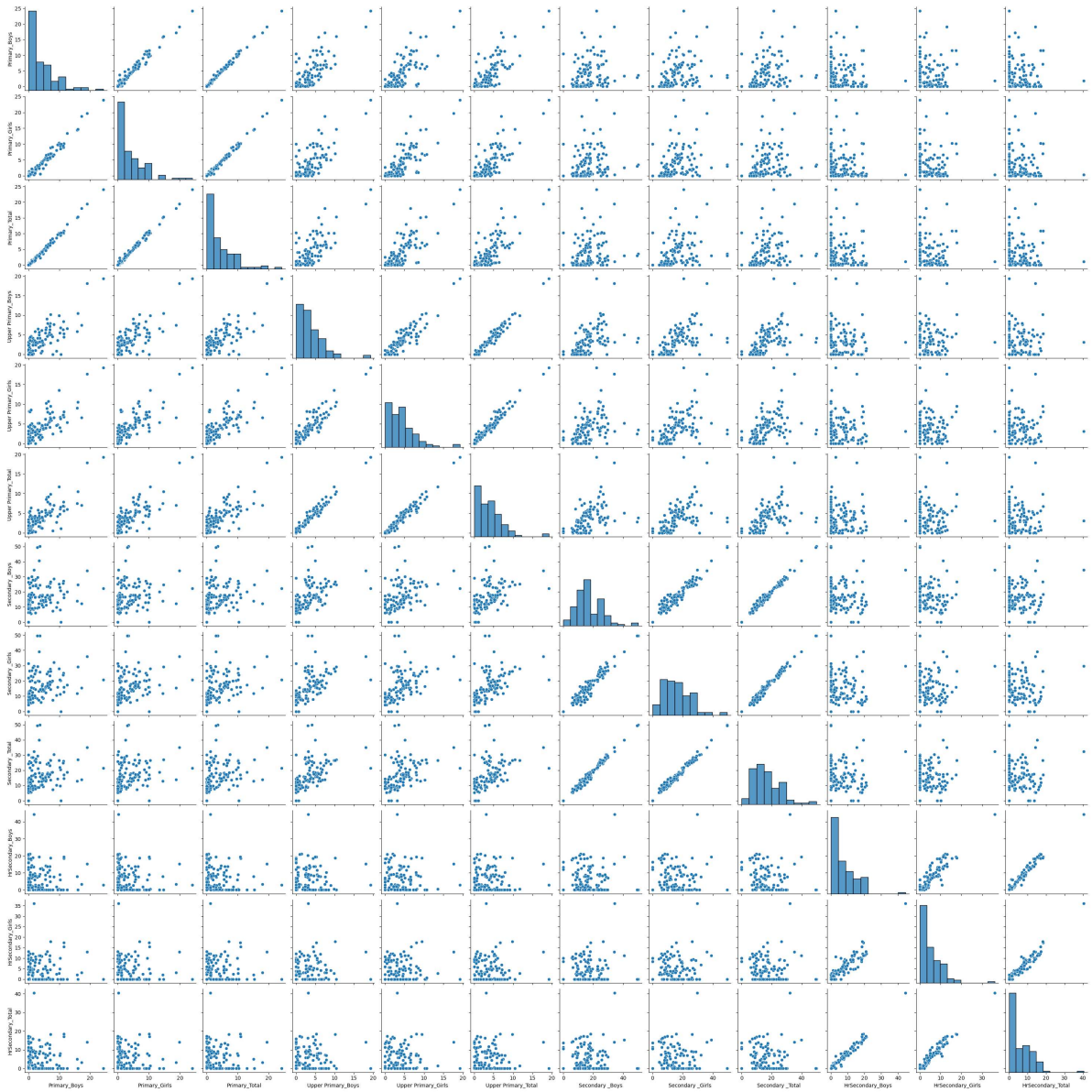
Out[18]:

	Primary_Boys	Primary_Girls	Primary_Total	Upper Primary_Boys	Upper Primary_Girls	Upper Primary_Total
count	110.000000	110.000000	110.000000	110.000000	110.000000	110.000000
mean	4.293455	4.010818	4.150000	3.581909	4.169455	3.839000
std	4.674719	4.553512	4.601164	3.388699	3.444964	3.351027
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.540000	0.590000	0.665000	0.782500	1.580000	1.200000
50%	2.900000	2.440000	2.885000	3.120000	3.535000	3.370000
75%	6.742500	5.870000	6.300000	5.422500	5.835000	5.462500
max	24.270000	23.930000	24.110000	19.350000	19.210000	19.280000

Exploratory Data Analysis

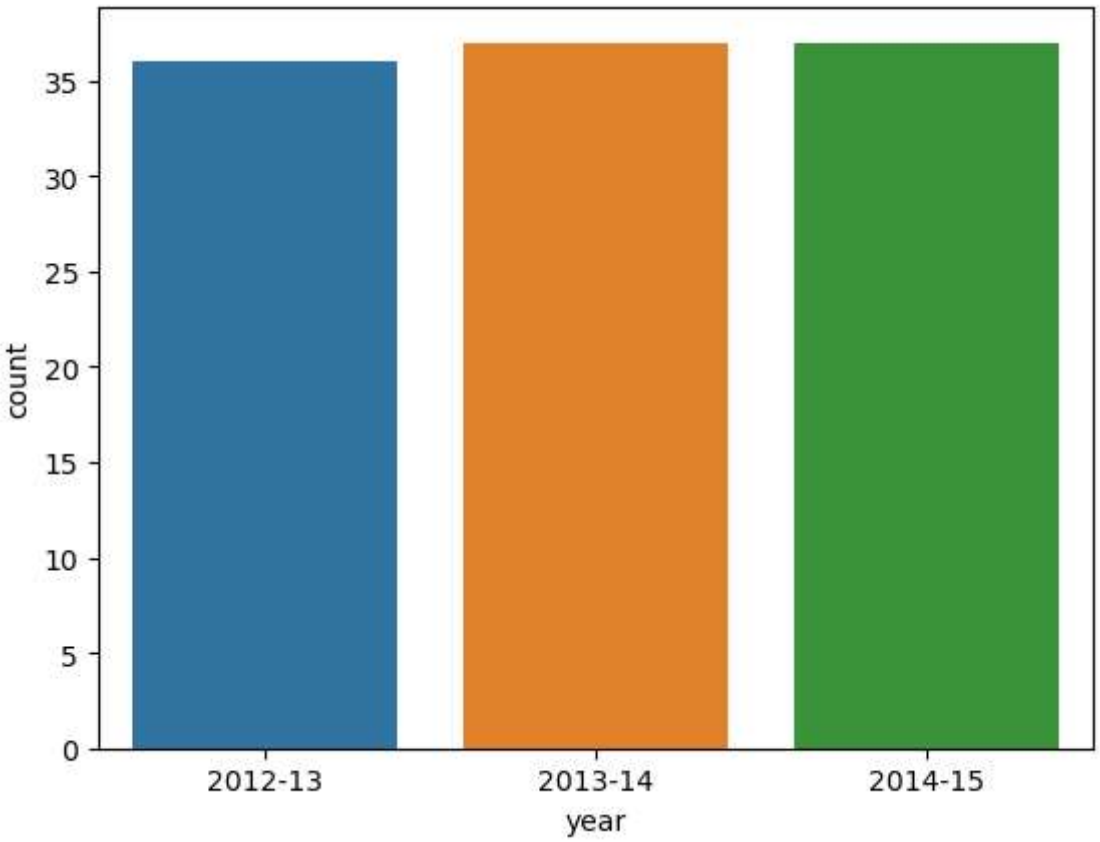
In [19]:
```
sns.pairplot(data=dropout_df)
```

Out[19]:
```
<seaborn.axisgrid.PairGrid at 0x190ad4926d0>
```



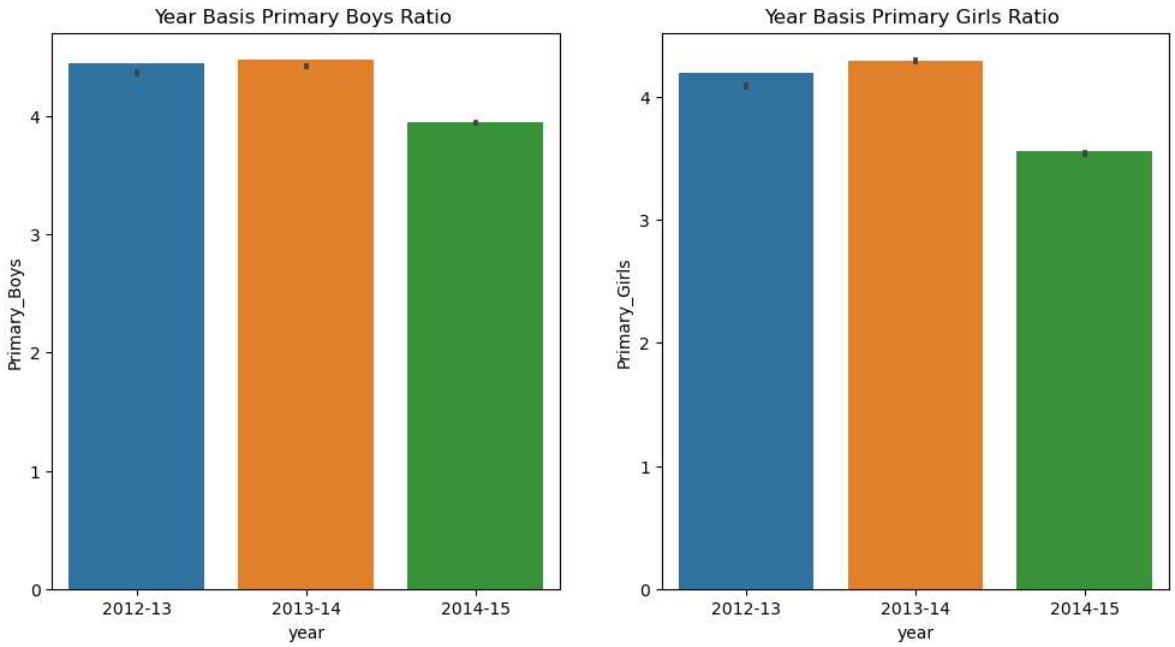
In [20]:
```
sns.countplot(x='year',data=dropout_df)
```

Out[20]:
```
<AxesSubplot:xlabel='year', ylabel='count'>
```



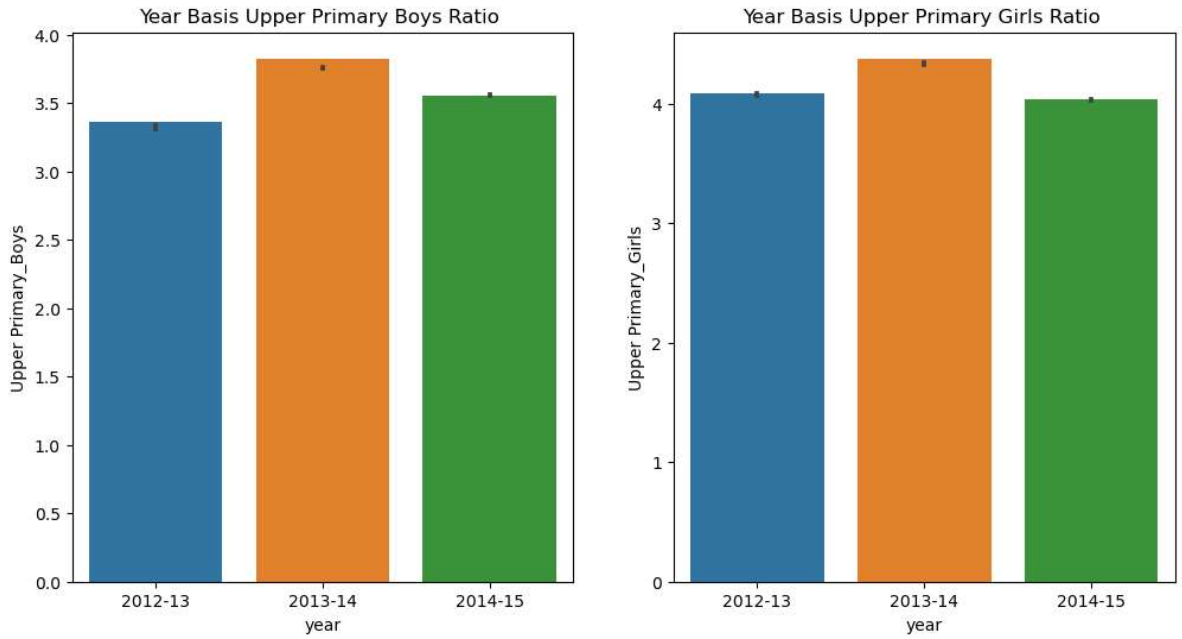
In [21]:
```
plt.figure(figsize=(12,6))

plt.subplot(1,2,1)
sns.barplot(dropout_df['year'],dropout_df['Primary_Boys'],ci=True)
plt.title('Year Basis Primary Boys Ratio ')

plt.subplot(1,2,2)
sns.barplot(dropout_df['year'],dropout_df['Primary_Girls'],ci=True)
plt.title('Year Basis Primary Girls Ratio ')
```

Out[21]:
```
Text(0.5, 1.0, 'Year Basis Primary Girls Ratio ')
```



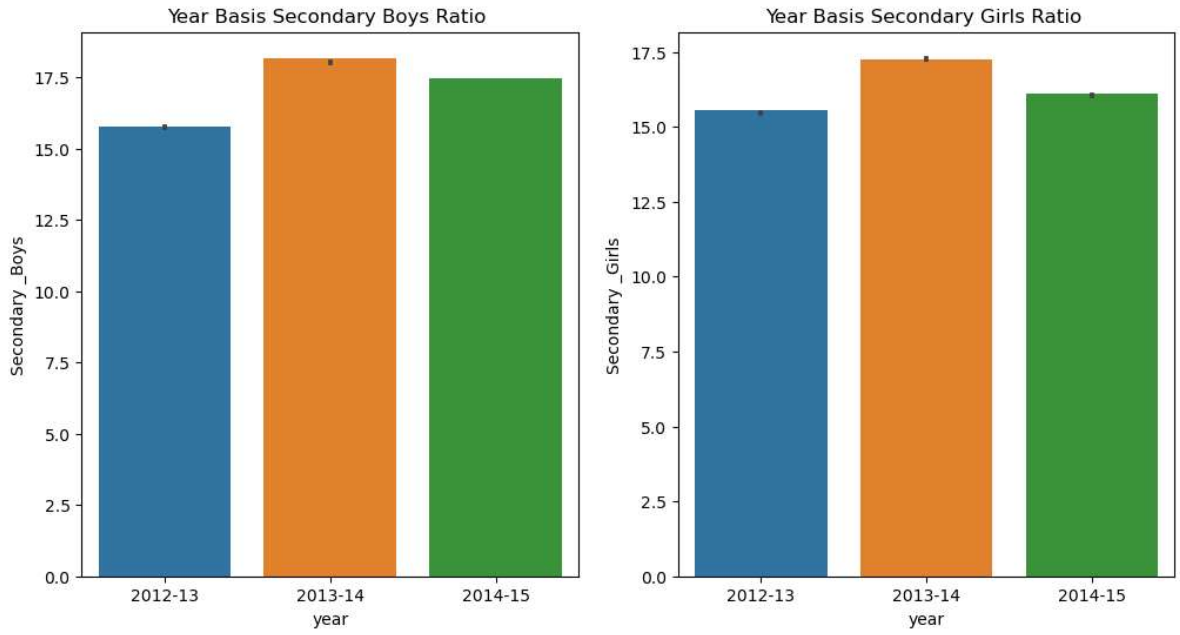
In [22]:
```
plt.figure(figsize=(12,6))

plt.subplot(1,2,1)
sns.barplot(dropout_df['year'],dropout_df['Upper Primary_Boys'],ci=True)
plt.title('Year Basis Upper Primary Boys Ratio ')

plt.subplot(1,2,2)
sns.barplot(dropout_df['year'],dropout_df['Upper Primary_Girls'],ci=True)
plt.title('Year Basis Upper Primary Girls Ratio ')
```

Out[22]: Text(0.5, 1.0, 'Year Basis Upper Primary Girls Ratio ')



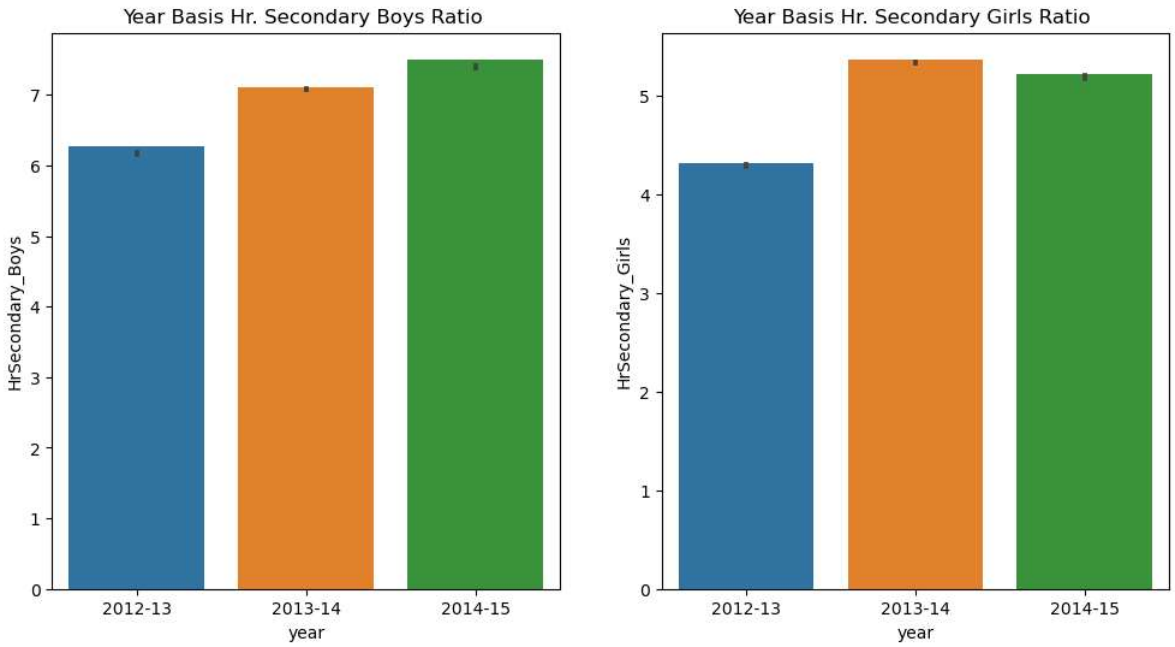
In [23]:
```
plt.figure(figsize=(12,6))

plt.subplot(1,2,1)
sns.barplot(dropout_df['year'],dropout_df['Secondary _Boys'],ci=True)
plt.title('Year Basis Secondary Boys Ratio ')

plt.subplot(1,2,2)
sns.barplot(dropout_df['year'],dropout_df['Secondary _Girls'],ci=True)
plt.title('Year Basis Secondary Girls Ratio ')
```

Out[23]:
```
Text(0.5, 1.0, 'Year Basis Secondary Girls Ratio ')
```

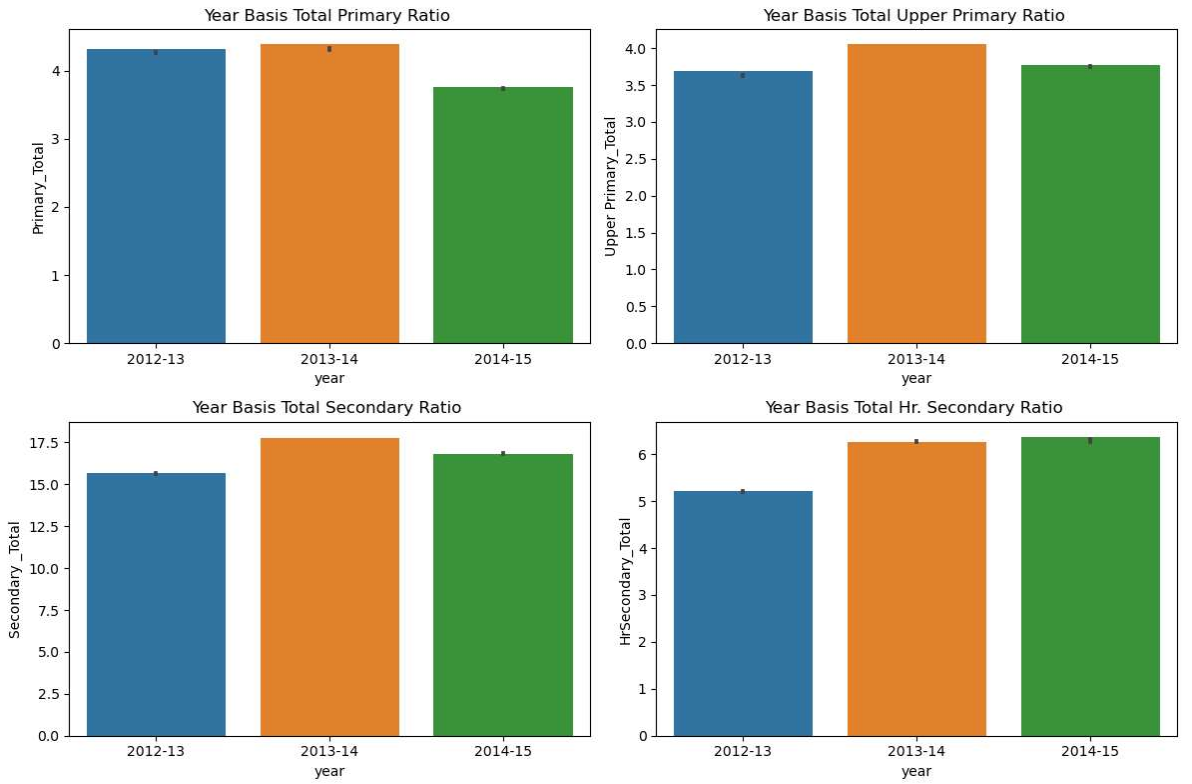


In [24]:
```
plt.figure(figsize=(12,6))

plt.subplot(1,2,1)
sns.barplot(dropout_df['year'],dropout_df['HrSecondary_Boys'],ci=True)
plt.title('Year Basis Hr. Secondary Boys Ratio ')

plt.subplot(1,2,2)
sns.barplot(dropout_df['year'],dropout_df['HrSecondary_Girls'],ci=True)
plt.title('Year Basis Hr. Secondary Girls Ratio ')
```

Out[24]:
```
Text(0.5, 1.0, 'Year Basis Hr. Secondary Girls Ratio ')
```



In [25]:
```
plt.figure(figsize=(12,8))

plt.subplot(2,2,1)
sns.barplot(dropout_df['year'],dropout_df['Primary_Total'],ci=True)
plt.title('Year Basis Total Primary Ratio ')

plt.subplot(2,2,2)
sns.barplot(dropout_df['year'],dropout_df['Upper Primary_Total'],ci=True)
plt.title('Year Basis Total Upper Primary Ratio ')

plt.subplot(2,2,3)
sns.barplot(dropout_df['year'],dropout_df['Secondary _Total'],ci=True)
plt.title('Year Basis Total Secondary Ratio ')

plt.subplot(2,2,4)
sns.barplot(dropout_df['year'],dropout_df['HrSecondary_Total'],ci=True)
plt.title('Year Basis Total Hr. Secondary Ratio ')

plt.tight_layout()
```



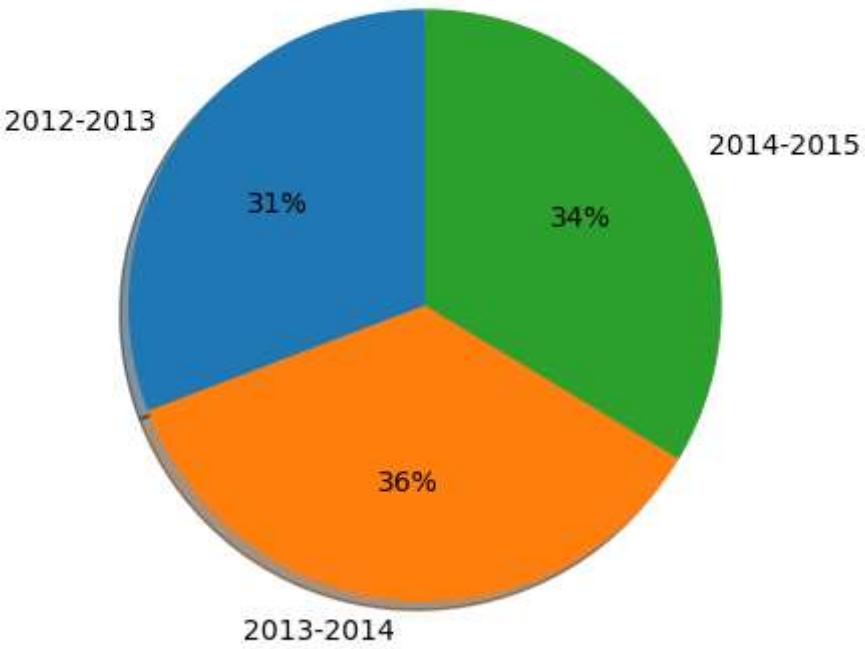
In [26]:
```
year_dr_1 = dropout_df[['Primary_Total','Upper Primary_Total','Secondary _Tota

year_dr_2 = dropout_df[['Primary_Total','Upper Primary_Total','Secondary _Tota

year_dr_3 = dropout_df[['Primary_Total','Upper Primary_Total','Secondary _Tota
```

In [27]:
```
print(year_dr_1.sum())
print(year_dr_2.sum())
print(year_dr_3.sum())
```

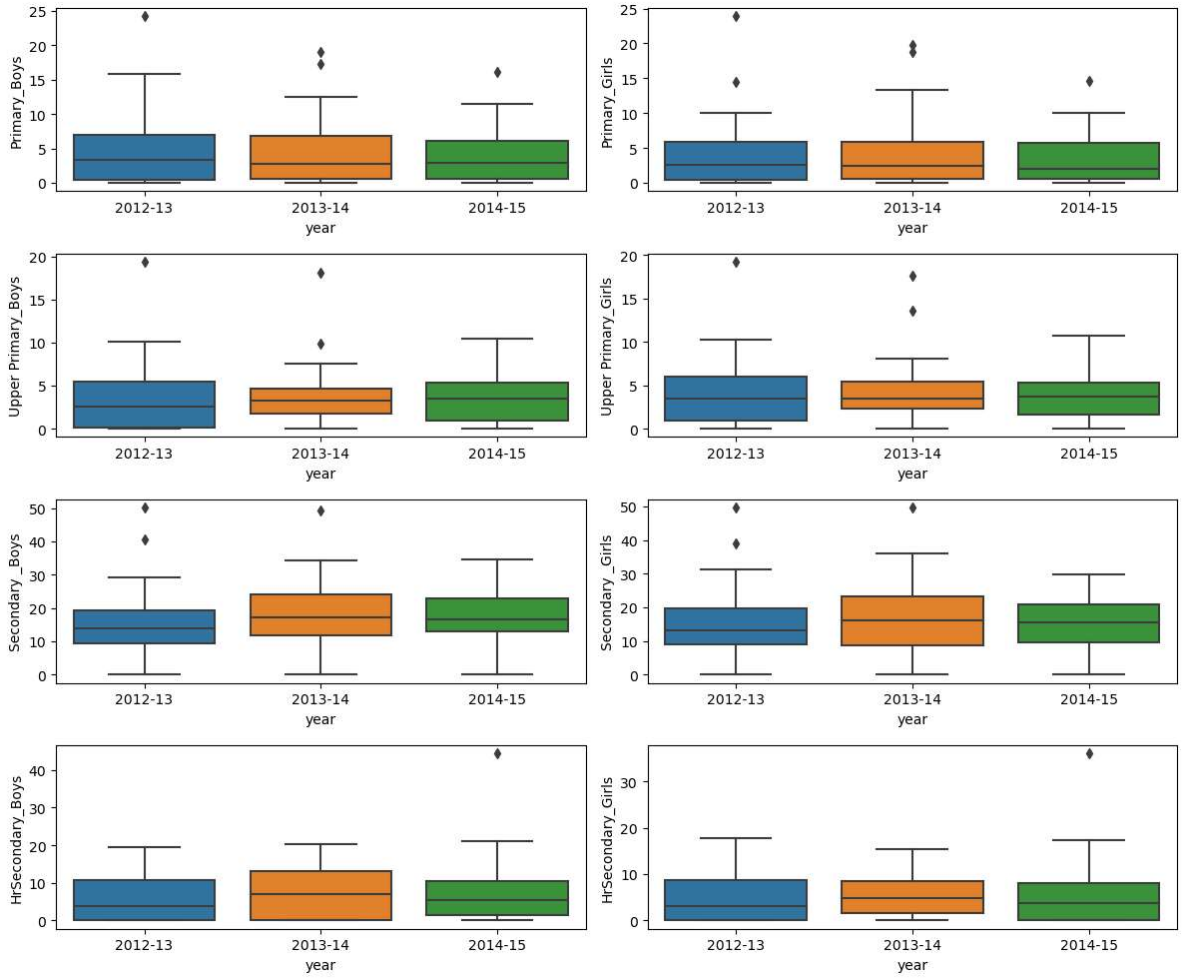
```
1040.51
1200.4200000000003
1137.02
```

In [28]:
```
years_duration_data = [year_dr_1.sum(), year_dr_2.sum(), year_dr_3.sum()]
years_duration = ['2012-2013', '2013-2014', '2014-2015']

plt.pie(years_duration_data, labels=years_duration, autopct='%.0f%%',shadow=Tr
```

Out[28]:
```
([<matplotlib.patches.Wedge at 0x190b7e0a730>,
  <matplotlib.patches.Wedge at 0x190b7e1c130>,
  <matplotlib.patches.Wedge at 0x190b7e1cac0>],
 [Text(-0.9059446817599497, 0.6239104371549362, '2012-2013'),
  Text(-0.09860046441169701, -1.0955719731801274, '2013-2014'),
  Text(0.9582231973424828, 0.5401928397107365, '2014-2015')],
 [Text(-0.49415164459633615, 0.3403147839026925, '31%'),
  Text(-0.05378207149728927, -0.5975847126437058, '36%'),
  Text(0.5226671985504451, 0.2946506398422199, '34%')])
```



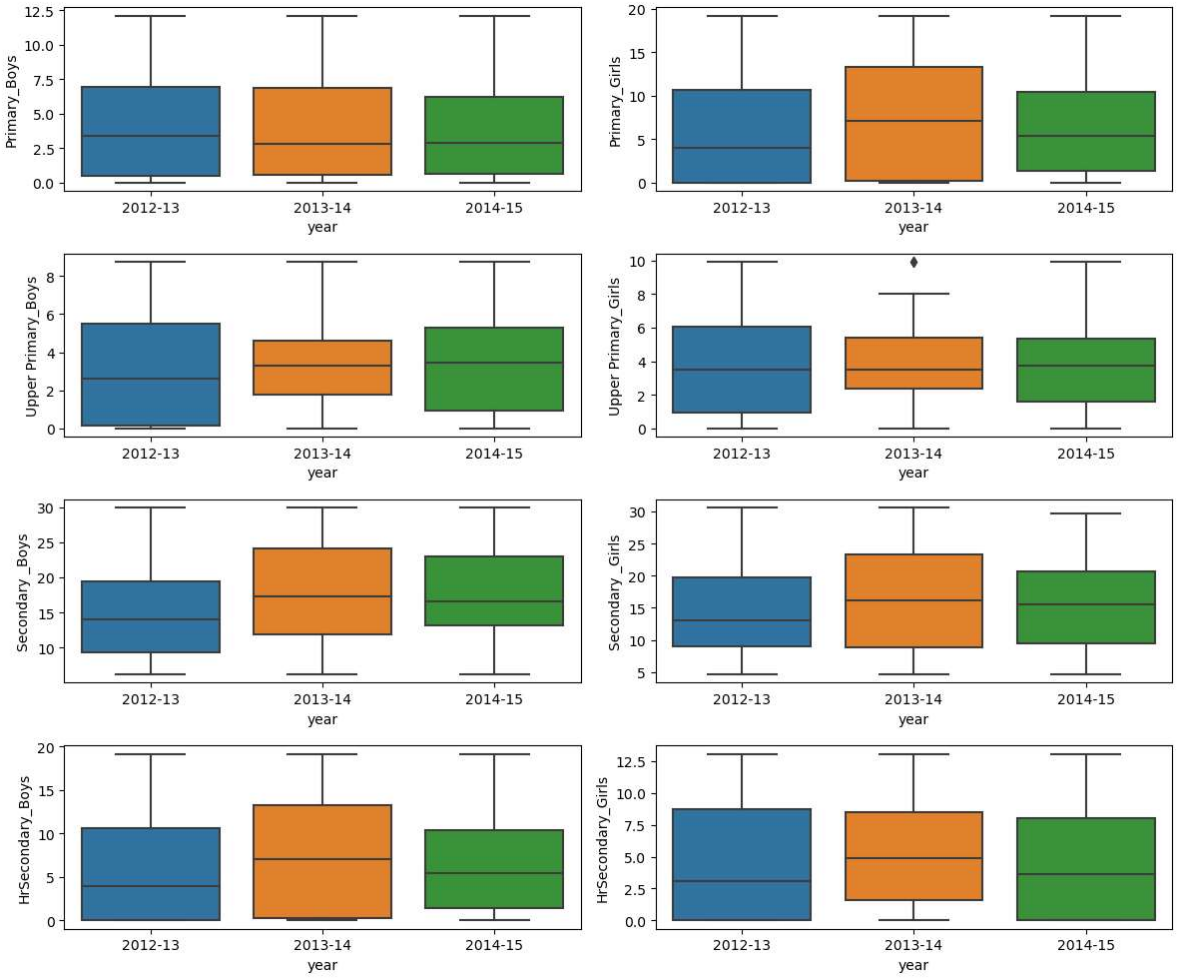
In [29]:
```
plt.figure(figsize=(12,10))

plt.subplot(4,2,1)
sns.boxplot(x='year',y='Primary_Boys',data=dropout_df)

plt.subplot(4,2,2)
sns.boxplot(x='year',y='Primary_Girls',data=dropout_df)

plt.subplot(4,2,3)
sns.boxplot(x='year',y='Upper Primary_Boys',data=dropout_df)

plt.subplot(4,2,4)
sns.boxplot(x='year',y='Upper Primary_Girls',data=dropout_df)

plt.subplot(4,2,5)
sns.boxplot(x='year',y='Secondary _Boys',data=dropout_df)

plt.subplot(4,2,6)
sns.boxplot(x='year',y='Secondary _Girls',data=dropout_df)

plt.subplot(4,2,7)
sns.boxplot(x='year',y='HrSecondary_Boys',data=dropout_df)

plt.subplot(4,2,8)
sns.boxplot(x='year',y='HrSecondary_Girls',data=dropout_df)

plt.tight_layout()
```



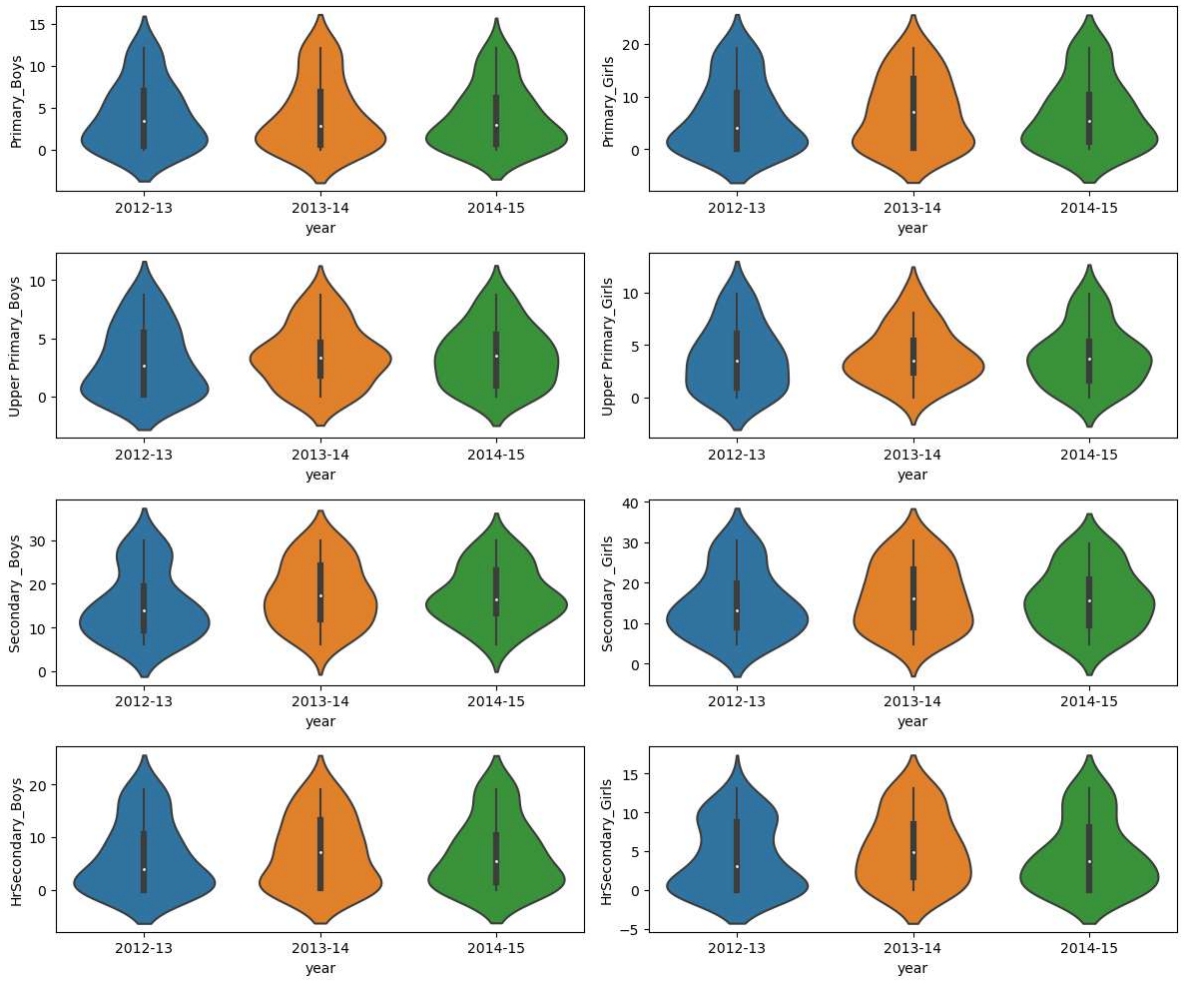
Outliers Removal

In [30]:
```
dropout_df['Primary_Boys']=dropout_df['Primary_Boys'].clip(lower=dropout_df['P
dropout_df['Primary_Girls']=dropout_df['HrSecondary_Boys'].clip(lower=dropout_

dropout_df['Upper Primary_Boys']=dropout_df['Upper Primary_Boys'].clip(lower=d
dropout_df['Upper Primary_Girls']=dropout_df['Upper Primary_Girls'].clip(lower

dropout_df['Secondary _Boys']=dropout_df['Secondary _Boys'].clip(lower=dropout
dropout_df['Secondary _Girls']=dropout_df['Secondary _Girls'].clip(lower=dropc

dropout_df['HrSecondary_Boys']=dropout_df['HrSecondary_Boys'].clip(lower=dropc
dropout_df['HrSecondary_Girls']=dropout_df['HrSecondary_Girls'].clip(lower=drc
```

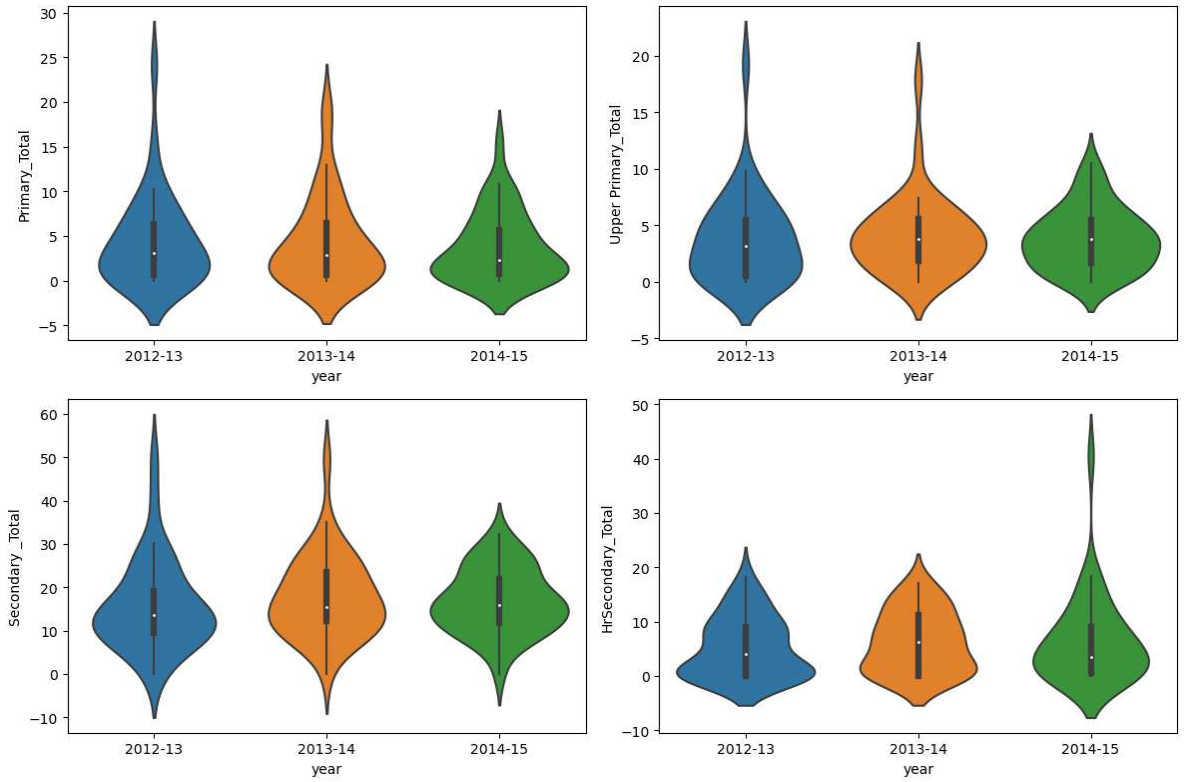
In [31]:
```
plt.figure(figsize=(12,10))

plt.subplot(4,2,1)
sns.boxplot(x='year',y='Primary_Boys',data=dropout_df)

plt.subplot(4,2,2)
sns.boxplot(x='year',y='Primary_Girls',data=dropout_df)

plt.subplot(4,2,3)
sns.boxplot(x='year',y='Upper Primary_Boys',data=dropout_df)

plt.subplot(4,2,4)
sns.boxplot(x='year',y='Upper Primary_Girls',data=dropout_df)

plt.subplot(4,2,5)
sns.boxplot(x='year',y='Secondary _Boys',data=dropout_df)

plt.subplot(4,2,6)
sns.boxplot(x='year',y='Secondary _Girls',data=dropout_df)

plt.subplot(4,2,7)
sns.boxplot(x='year',y='HrSecondary_Boys',data=dropout_df)

plt.subplot(4,2,8)
sns.boxplot(x='year',y='HrSecondary_Girls',data=dropout_df)

plt.tight_layout()
```


Primary_Boys
Primary_Girls
Upper Primary_Boys
Upper Primary_Girls
Secondary _Boys
Secondary _Girls
HrSecondary_Boys
HrSecondary_Girls
2012-13
2013-14
2014-15
year

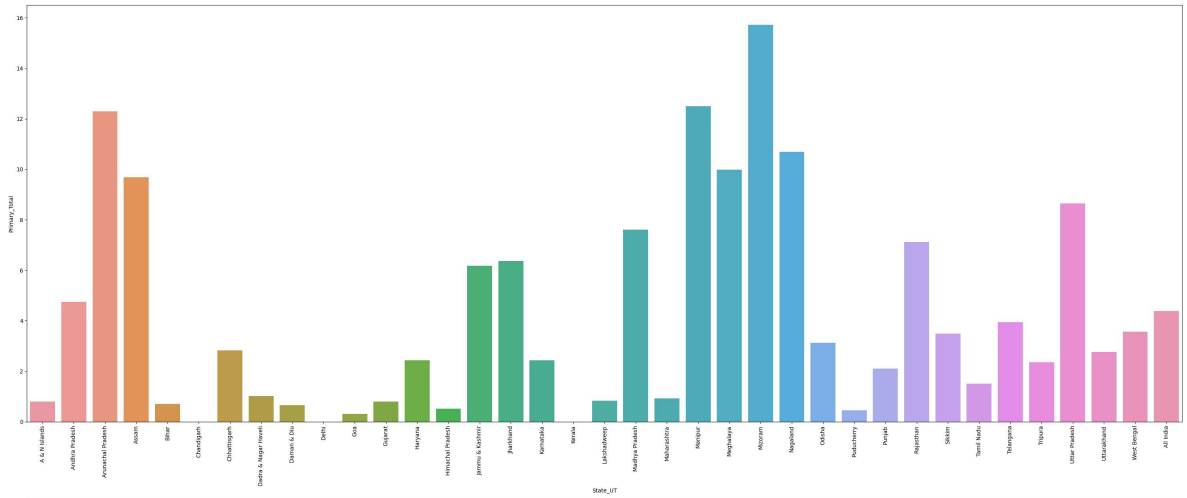
In [32]:
```
plt.figure(figsize=(12,10))

plt.subplot(4,2,1)
sns.violinplot(x='year',y='Primary_Boys',data=dropout_df)

plt.subplot(4,2,2)
sns.violinplot(x='year',y='Primary_Girls',data=dropout_df)

plt.subplot(4,2,3)
sns.violinplot(x='year',y='Upper Primary_Boys',data=dropout_df)

plt.subplot(4,2,4)
sns.violinplot(x='year',y='Upper Primary_Girls',data=dropout_df)

plt.subplot(4,2,5)
sns.violinplot(x='year',y='Secondary _Boys',data=dropout_df)

plt.subplot(4,2,6)
sns.violinplot(x='year',y='Secondary _Girls',data=dropout_df)

plt.subplot(4,2,7)
sns.violinplot(x='year',y='HrSecondary_Boys',data=dropout_df)

plt.subplot(4,2,8)
sns.violinplot(x='year',y='HrSecondary_Girls',data=dropout_df)

plt.tight_layout()
```

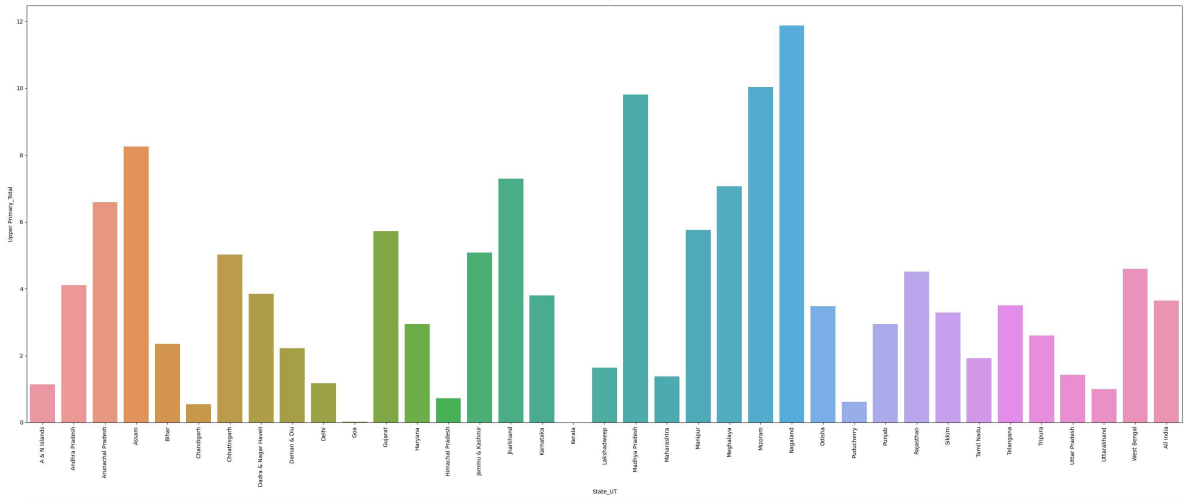

Primary_Boys
Primary_Girls
Upper Primary_Boys
Upper Primary_Girls
Secondary _Boys
Secondary _Girls
HrSecondary_Boys
HrSecondary_Girls
2012-13
2013-14
2014-15
year

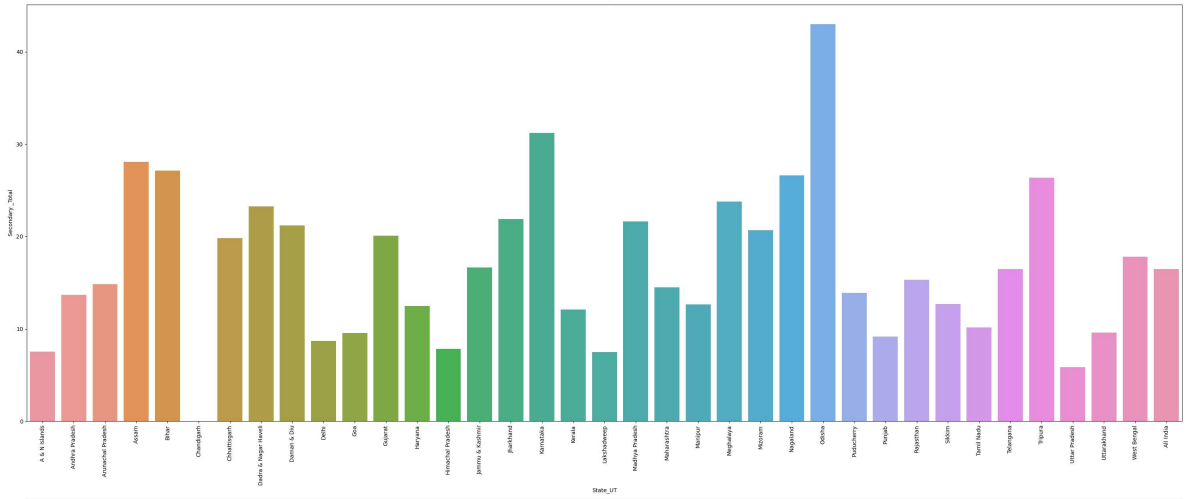
In [33]:
```
plt.figure(figsize=(12,8))

plt.subplot(2,2,1)
sns.violinplot(x=dropout_df['year'],y=dropout_df['Primary_Total'])

plt.subplot(2,2,2)
sns.violinplot(x=dropout_df['year'],y=dropout_df['Upper Primary_Total'])

plt.subplot(2,2,3)
sns.violinplot(x=dropout_df['year'],y=dropout_df['Secondary _Total'])

plt.subplot(2,2,4)
sns.violinplot(x=dropout_df['year'],y=dropout_df['HrSecondary_Total'])

plt.tight_layout()
```

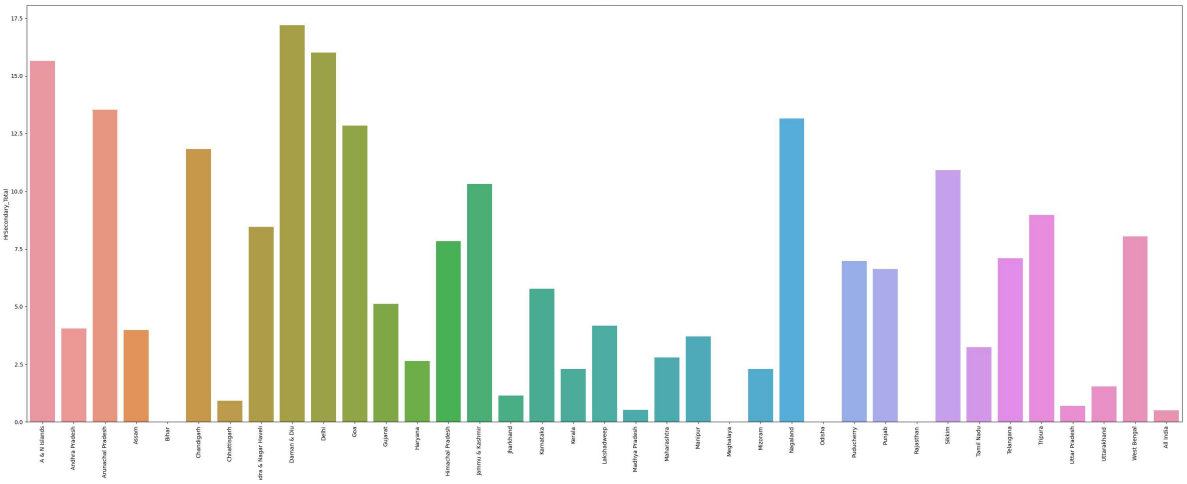


In [34]:
```
plt.figure(figsize=(30,50))
plt.subplot(4,1,1)
sns.barplot(x='State_UT',y='Primary_Total',data=dropout_df,ci=True)
plt.xticks(rotation = 90)

plt.subplot(4,1,2)
sns.barplot(x='State_UT',y='Upper Primary_Total',data=dropout_df,ci=True)
plt.xticks(rotation = 90)

plt.subplot(4,1,3)
sns.barplot(x='State_UT',y='Secondary _Total',data=dropout_df,ci=True)
plt.xticks(rotation = 90)

plt.subplot(4,1,4)
sns.barplot(x='State_UT',y='HrSecondary_Total',data=dropout_df,ci=True)
plt.xticks(rotation = 90)

plt.xticks(rotation = 90)
plt.tight_layout()
```


Primary_Total
State_UT
A & N Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra & Nagar Haveli
Daman & Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
All India


Upper Primary_Total
State_UT
A & N Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra & Nagar Haveli
Daman & Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
All India


Secondary_Total
State_UT
A & N Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra & Nagar Haveli
Daman & Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
All India


HrSecondary_Total
A & N Islands
Andhra Pradesh
Arunachal Pradesh
Assam
Bihar
Chandigarh
Chhattisgarh
Dadra & Nagar Haveli
Daman & Diu
Delhi
Goa
Gujarat
Haryana
Himachal Pradesh
Jammu & Kashmir
Jharkhand
Karnataka
Kerala
Lakshadweep
Madhya Pradesh
Maharashtra
Manipur
Meghalaya
Mizoram
Nagaland
Odisha
Puducherry
Punjab
Rajasthan
Sikkim
Tamil Nadu
Telangana
Tripura
Uttar Pradesh
Uttarakhand
West Bengal
All India

State_UT

Conclusion

1. According to above Analysis it came to know that most of the students take dropout in year 2013-2014 (36%) duration and after that in year 2014-2015 (34%) duration.
2. One more information is came to know from above Analysis is that, in the education stage (Primary and Upper Primary), 'Girls' or 'Female' students are take the most dropout. And in the education stage (Secondary and Hr. Secondary), 'Boys' or 'Male' students are take the most dropout.

In [ ]: