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

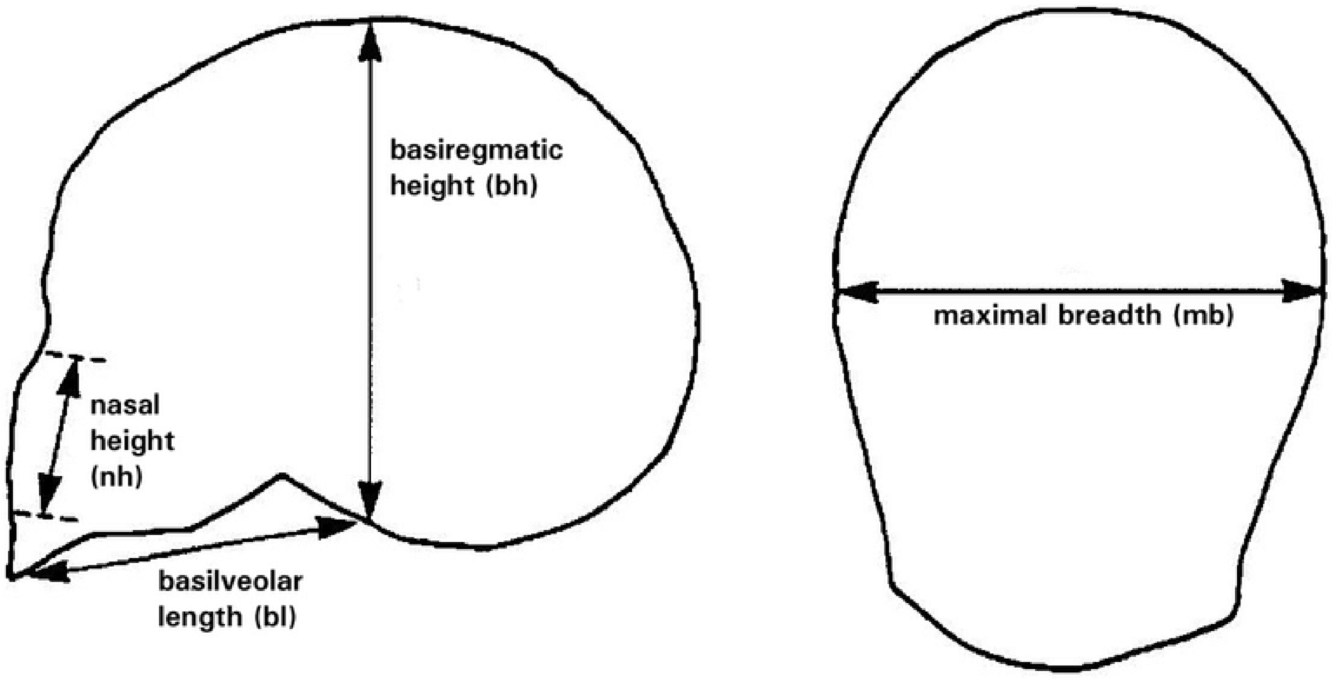
In [2]:
```
df = pd.read_csv('skull.txt',sep='\t')
df.rename(columns={
    'MB':'Maximal Breadth of Skull',
    'BH':'Basibregmatic Height of Skull',
    'BL':'Basialveolar Length of Skull',
    'NH':'Nasal Height of Skull'
},inplace = True)
```

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

Out[3]:

	Maximal Breadth of Skull	Basibregmatic Height of Skull	Basialveolar Length of Skull	Nasal Height of Skull	Year
0	131	138	89	49	-4000
1	125	131	92	48	-4000
2	131	132	99	50	-4000
3	119	132	96	44	-4000
4	136	143	100	54	-4000

Four measurements of male Egyptian skulls from 5 different time periods. Thirty skulls are measured from each time period.



Graphical description of the four skull mesurements

- Dataset is a text file, each observation on a new line with variables separated by tabs
- There are 150 skull measurments, each have four measurments in mm, numbers are integers
- Skulls are from five separate historical eras from the years -4000, -3300, -1850, -200, 150
- Each era has 30 Skulls
- There are no missing values
- Refereing to the skull diagrams, the four skull measurements are:
    1. MB - Maximal Breadth
    2. BH - Basiregmatic Hight, the hight of the skull
    3. BL - Basilveolar Length, the length of the skull
    4. NH - Nasal Height, as drawn in fig

Descriptive Statistics

In [4]:
```
# Before getting deep into the problem, let's try to get some descriptive statistics for numerical columns
df.describe().style.background_gradient(cmap = 'copper')
```

Out[4]:

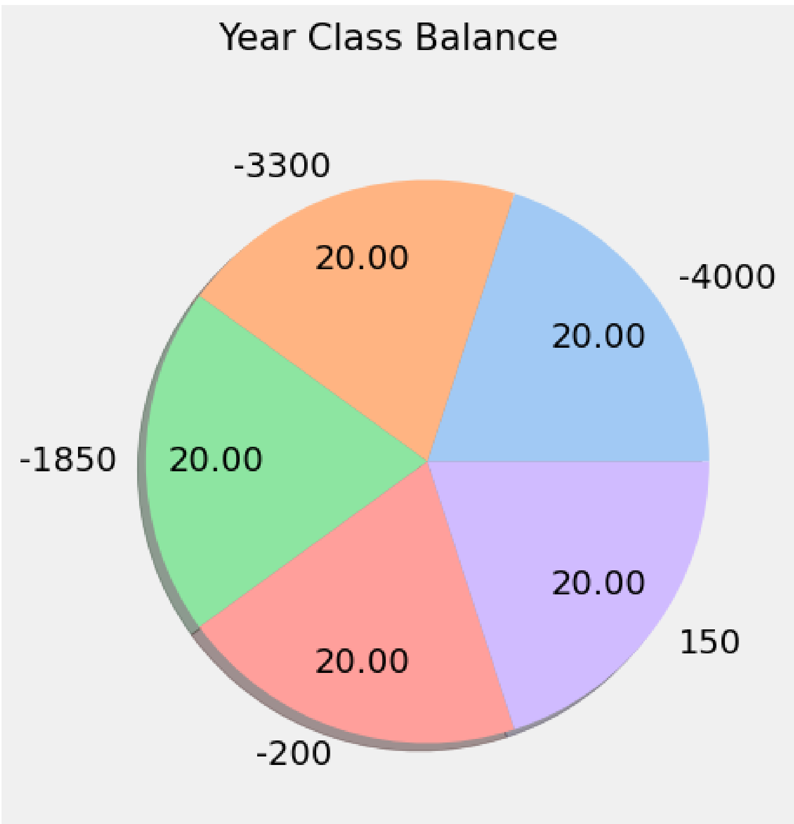
	Maximal Breadth of Skull	Basibregmatic Height of Skull	Basialveolar Length of Skull	Nasal Height of Skull	Year
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	133.973333	132.546667	96.460000	50.933333	-1840.000000
std	4.890680	4.939346	5.377844	3.207932	1645.432972
min	119.000000	120.000000	81.000000	44.000000	-4000.000000
25%	131.000000	129.000000	93.000000	49.000000	-3300.000000
50%	134.000000	133.000000	96.000000	51.000000	-1850.000000
75%	137.000000	136.000000	100.000000	53.000000	-200.000000
max	148.000000	145.000000	114.000000	60.000000	150.000000

In [5]:
```
df.groupby('Year').agg({
    'Maximal Breadth of Skull':'mean',
     'Basibregmatic Height of Skull':'mean',
     'Basialveolar Length of Skull':'mean',
     'Nasal Height of Skull':'mean',
}).style.background_gradient(cmap = 'copper')
```

Out[5]:

Year	Maximal Breadth of Skull	Basibregmatic Height of Skull	Basialveolar Length of Skull	Nasal Height of Skull
-4000	131.366667	133.600000	99.166667	50.533333
-3300	132.366667	132.700000	99.066667	50.233333
-1850	134.466667	133.800000	96.033333	50.566667
-200	135.500000	132.300000	94.533333	51.966667
150	136.166667	130.333333	93.500000	51.366667

In [6]:
```
# Because we want to determine how the skull size has changed over time,let's check the Target Class Balance

plt.rcParams['figure.figsize'] = (15, 5)
plt.style.use('fivethirtyeight')

plt.xlabel('Year', fontsize = 10)
df['Year'].value_counts().plot(kind = 'pie', autopct = '%.2f', startangle = 0,
                                     labels = ['-4000', '-3300', '-1850', '-200', '150'],
                                      shadow = True, pctdistance = 0.75,
                              colors=sns.color_palette('pastel')[0:5])

plt.axis('off')

plt.suptitle('Year Class Balance', fontsize = 15)
plt.show()
```



We can easily see that the Year Class is balanced, which means that the number of observations of each year is distributed uniformly.

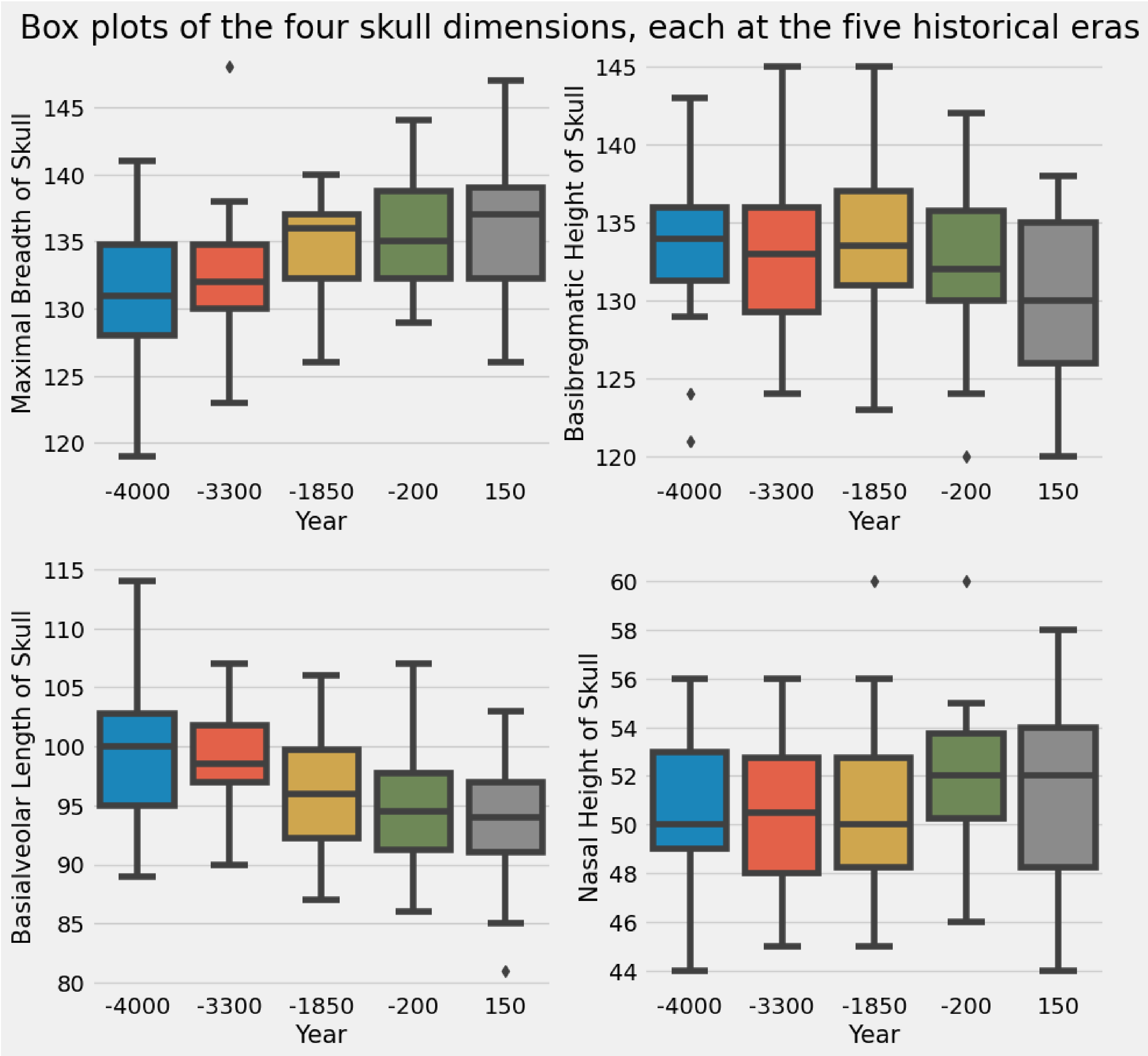
- Most of the times, when we use Machine Learning model with imbalanced classes, we have a very poor results which are competely biased towards the class having higher distribution.

Outlier Detection/ Anomaly Detection

The presence of outliers in a classification or regression dataset can result in a poor fit and lower predictive modeling performance. Instead, automatic outlier detection methods can be used in the modeling pipeline and compared, just like other data preparation transforms that may be applied to the dataset.

Some datasets can be imputed wrong which will not make any sense. Sometimes we can detect anomaly obsevation which we might consider to drop.

In [7]:
```
variables = list(df.columns)
rows=2
cols=2
f, axes = plt.subplots(rows, cols, figsize=(10,10))
f.suptitle('Box plots of the four skull dimensions, each at the five historical eras', fontsize=20, y=0.91)

for i in range(rows):
    for j in range(cols):
        b = sns.boxplot(  y=df[variables[i*rows+j]], x= df['Year'], data=df,  orient='v' , ax=axes[i,j])
        b.set_xlabel("Year",fontsize=15)
        b.set_ylabel(variables[i*rows+j],fontsize=15)
        variables = list(df.columns)
```



The shapes of the four skull dimension histograms is skewed and asymmetrical. This skewness suggest that the skull population contains several sources. Our Box plot of the four skull dimentions along the five historical eras shows that the skulls shapes had changed over the years. The average skull breadth increased over the years from 131 to 137 mm while the average skull heights and length decreaded from 134 to 130 and 100 to 95 respectivaly. The nasal height changes are less evident but also show increase when comparing the later to the earlier eras. The cross correlation heatmap show that the skulls breadth is slightly correlated with the nasal height and negatively correlated with the length. The height is slighly correlated with the length and the nasal height. Researchers claim that these changes in the skulls shape caused by migration of new populations into Egypt during history and interbreeding with the original Egyptians.

Here, the Box plot, helps us to analyze the middle 50 percentile of the data, and we can clearly check the minimum, maximum, median, and outlier values.

We also check the Distribution of these attributes after checking the Box Plot so that we can be more clear about the Values present in these columns.

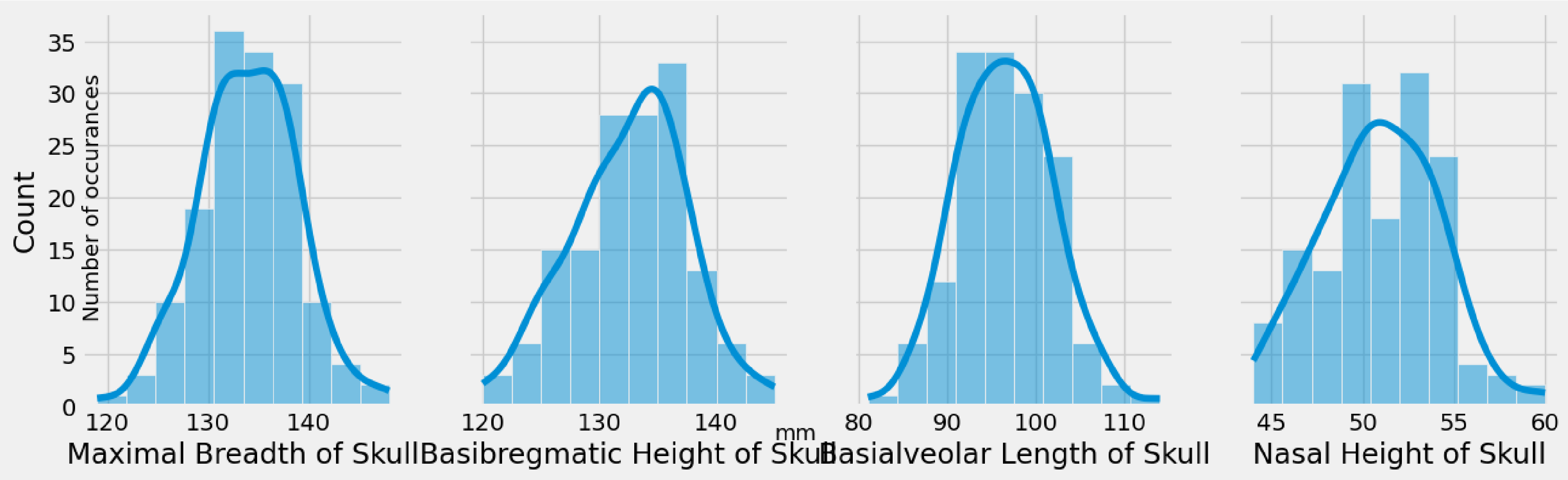
Univariate Analysis

In [8]:
```
# Check the length unqiue values of each column
for column in list(df.columns):
    print(column,':' ,len(set(list(df[column]))))
```

```
Maximal Breadth of Skull : 26
Basibregmatic Height of Skull : 24
Basialveolar Length of Skull : 27
Nasal Height of Skull : 16
Year : 5
```

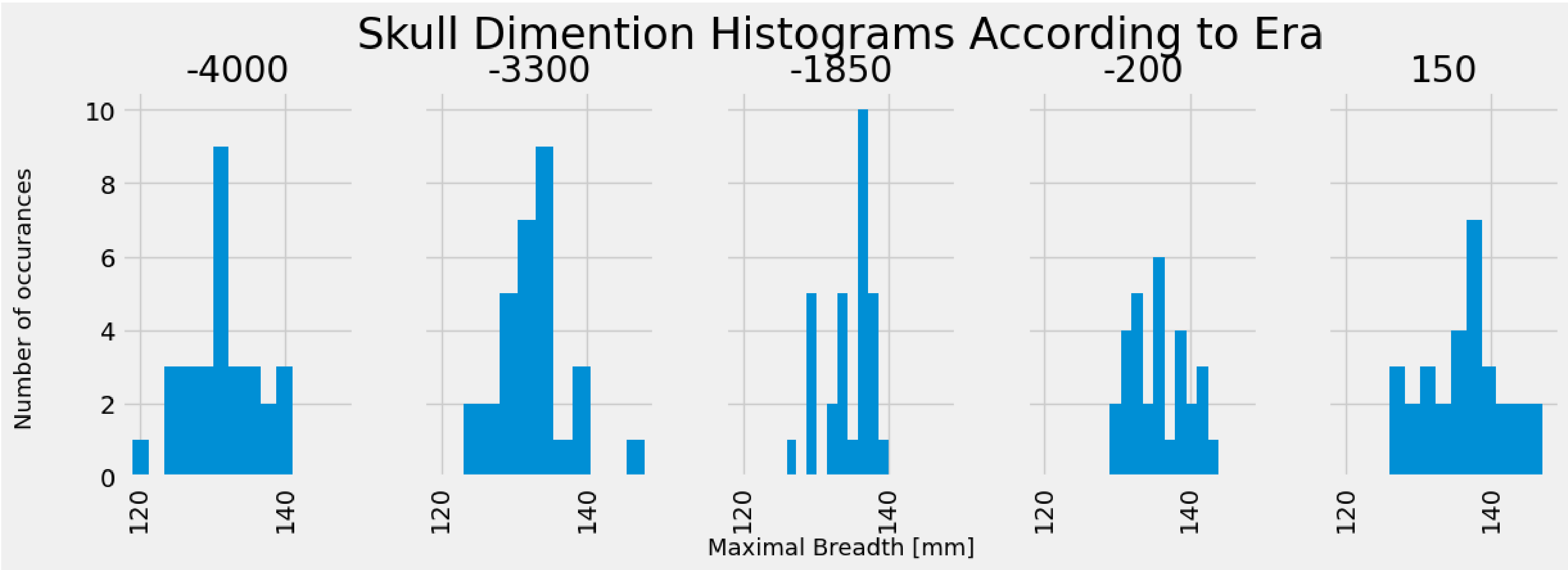
In [9]:
```
# set up figure & axes
fig, axs = plt.subplots(ncols=4,figsize=(14,4), sharex=False, sharey=True)

sns.histplot(data=df, x="Maximal Breadth of Skull",bins=10, kde=True,ax=axs[0])
sns.histplot(data=df, x="Basibregmatic Height of Skull",bins=10, kde=True,ax=axs[1])
sns.histplot(data=df, x="Basialveolar Length of Skull",bins=10, kde=True,ax=axs[2])
sns.histplot(data=df, x="Nasal Height of Skull",bins=10, kde=True,ax=axs[3])

fig.text(0.08, 0.5, 'Number of occurances', va='center', rotation='vertical', fontsize=13)
fig.text(0.5, 0.0, 'mm', ha='center', fontsize=13)
```

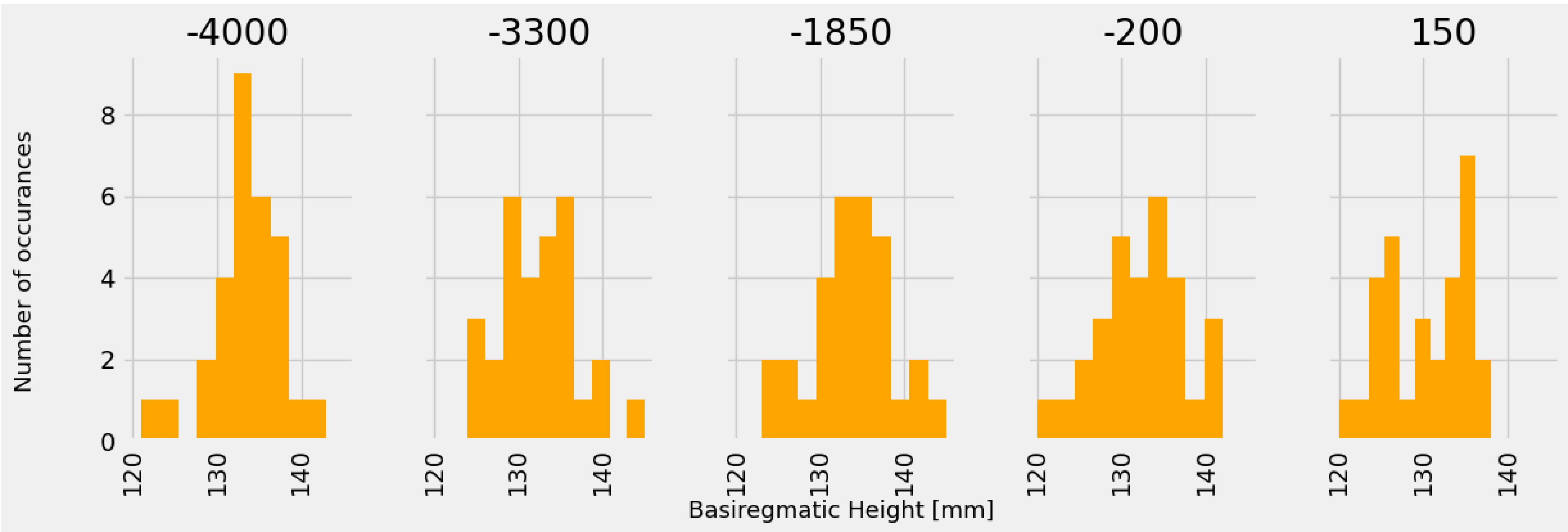
Out[9]: Text(0.5, 0.0, 'mm')

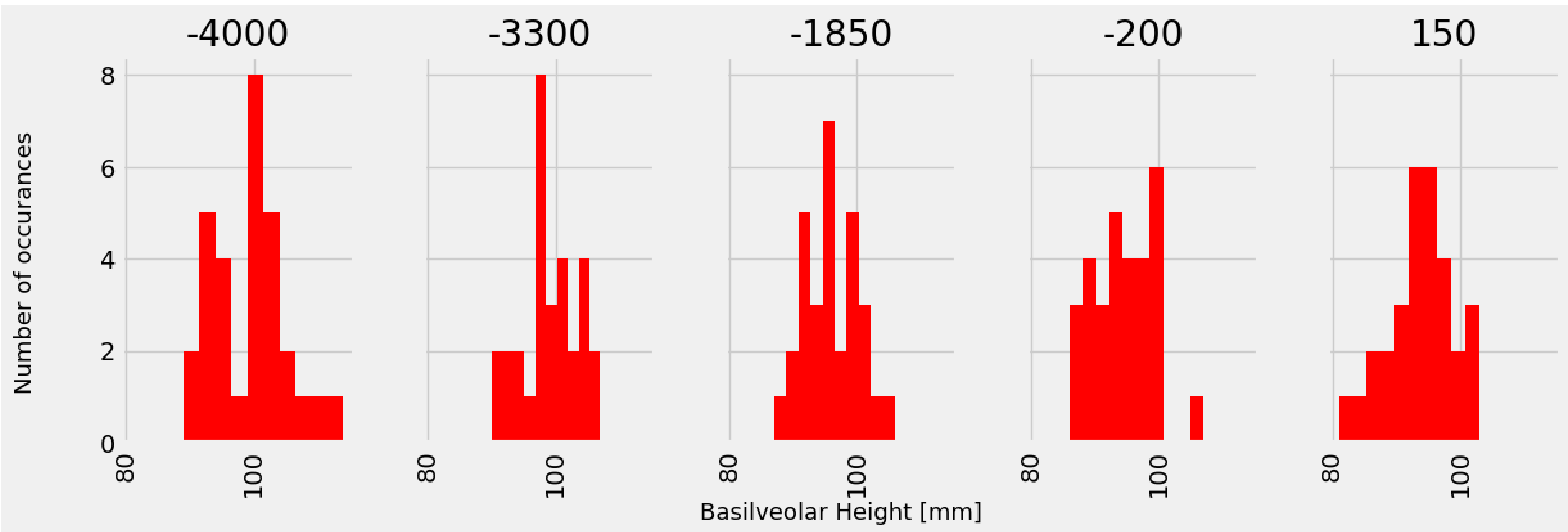


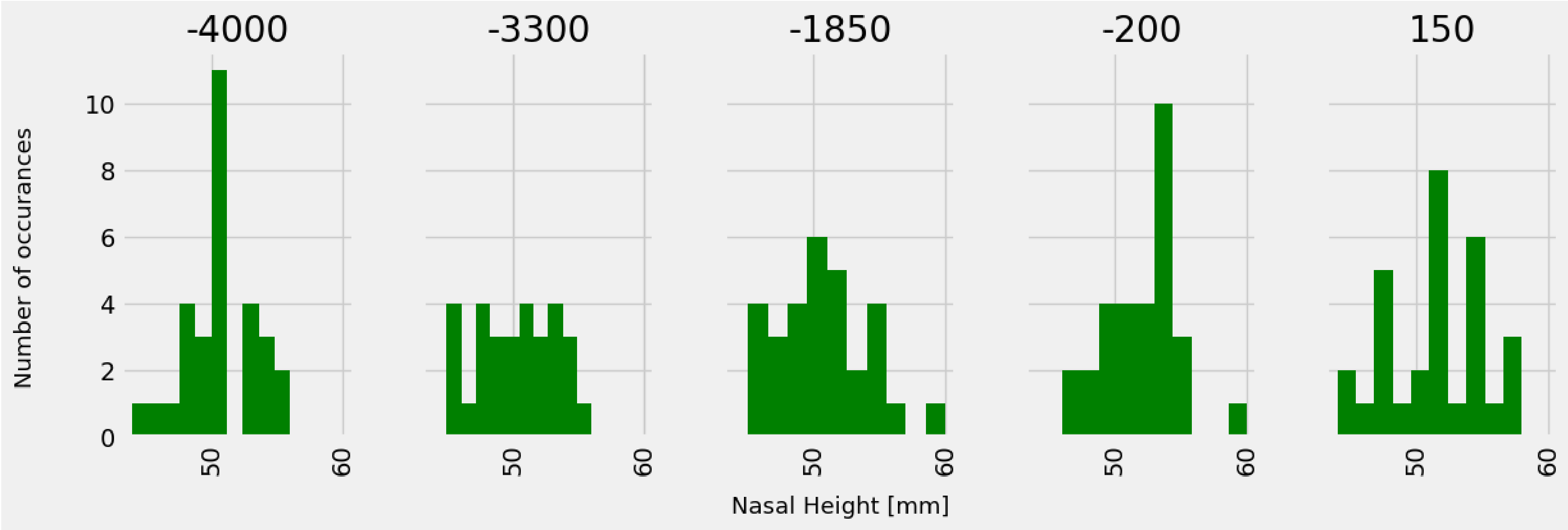
In [10]:
```
# set up figure & axes
fig, axes = plt.subplots(nrows=1, ncols=5, figsize=(14,4), sharex=True, sharey=True)

# drop sharex, sharey, layout & add ax=axes
#df1.hist(column='human_den',by='region', ax=axes)
df.hist(column='Maximal Breadth of Skull', by='Year', ax=axes)
# set title and axis labels
plt.suptitle('Skull Dimention Histograms According to Era', x=0.5, y=1.05, ha='center', fontsize='xx-large')
fig.text(0.5, 0.0, 'Maximal Breadth [mm]', ha='center', fontsize=13)
fig.text(0.04, 0.5, 'Number of occurances', va='center', rotation='vertical', fontsize=13)

fig, axes = plt.subplots(nrows=1, ncols=5, figsize=(14,4), sharex=True, sharey=True)
df.hist(column='Basibregmatic Height of Skull', by='Year', ax=axes, color='orange')
# set title and axis labels
#plt.suptitle('Maximal Breadth Histograms According to Era', x=0.5, y=1.05, ha='center', fontsize='xx-large')
fig.text(0.5, 0.0, 'Basiregmatic Height [mm]', ha='center', fontsize=13)
fig.text(0.04, 0.5, 'Number of occurances', va='center', rotation='vertical', fontsize=13)

fig, axes = plt.subplots(nrows=1, ncols=5, figsize=(14,4), sharex=True, sharey=True)
df.hist(column='Basialveolar Length of Skull', by='Year', ax=axes, color='red')
# set title and axis labels
#plt.suptitle('Maximal Breadth Histograms According to Era', x=0.5, y=1.05, ha='center', fontsize='xx-large')
fig.text(0.5, 0.0, 'Basilveolar Height [mm]', ha='center', fontsize=13)
fig.text(0.04, 0.5, 'Number of occurances', va='center', rotation='vertical', fontsize=13)

fig, axes = plt.subplots(nrows=1, ncols=5, figsize=(14,4), sharex=True, sharey=True)
df.hist(column='Nasal Height of Skull', by='Year', ax=axes, color='green')
# set title and axis labels
#plt.suptitle('Maximal Breadth Histograms According to Era', x=0.5, y=1.05, ha='center', fontsize='xx-large')
fig.text(0.5, 0.0, 'Nasal Height [mm]', ha='center', fontsize=13)
fig.text(0.04, 0.5, 'Number of occurances', va='center', rotation='vertical', fontsize=13)
```

Out[10]: `Text(0.04, 0.5, 'Number of occurances')`



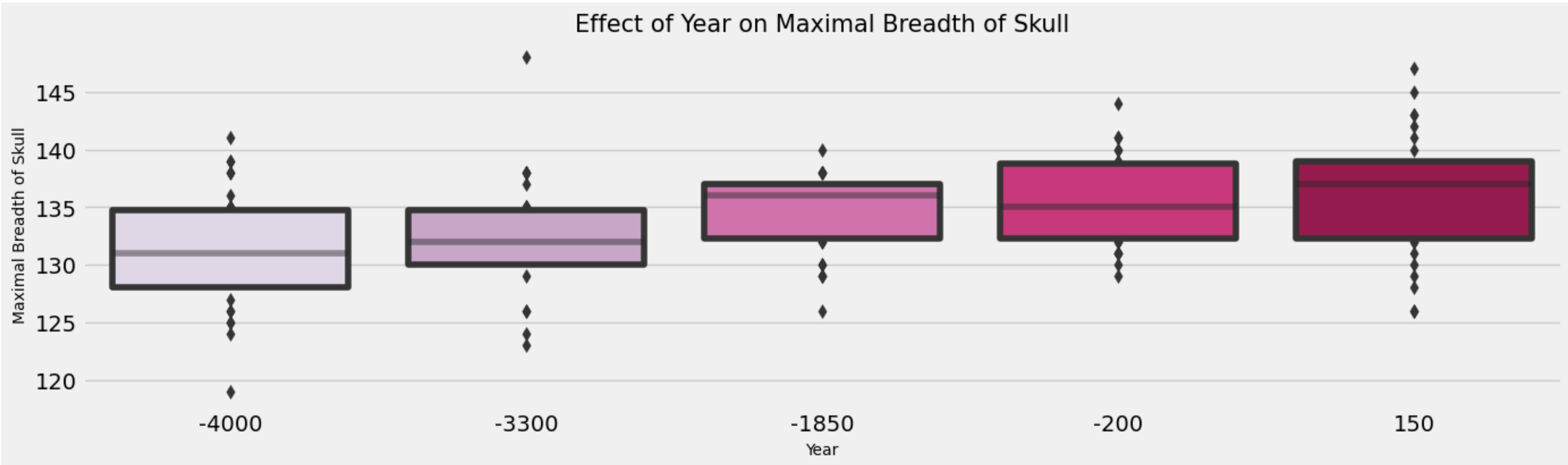






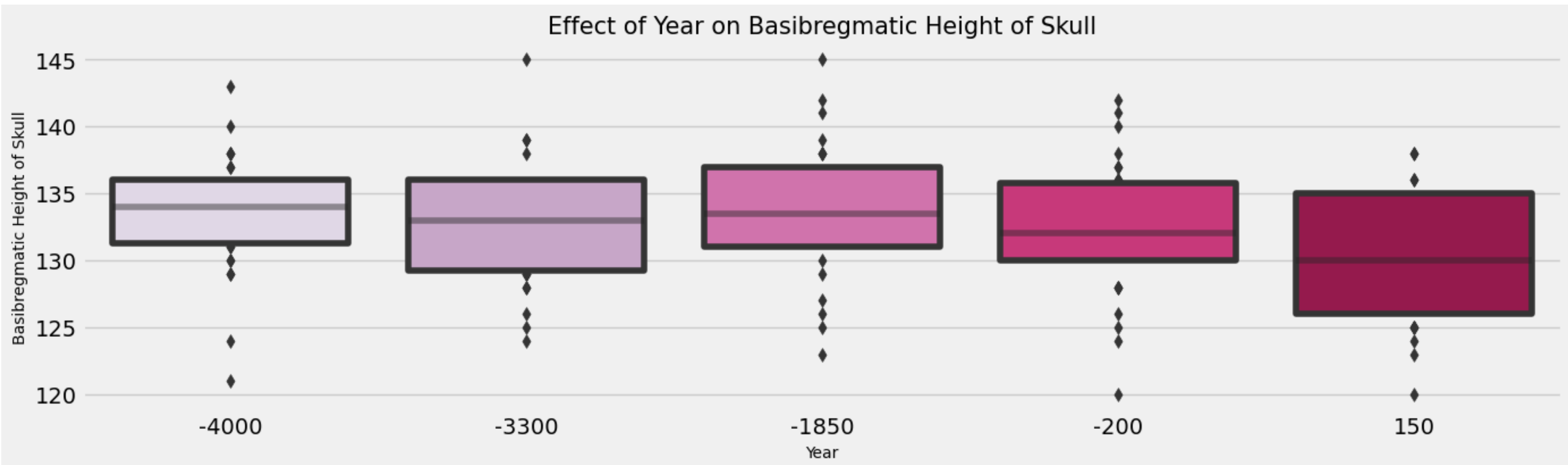
Bivariate Analysis & Multivariate Analysis

In [11]:
```
# Effect of Year on Maximal Breadth of Skull

plt.rcParams['figure.figsize'] = (15,4)
sns.boxenplot(x = df['Year'].values,y= df['Maximal Breadth of Skull'].values, palette = 'PuRd')
plt.title('Effect of Year on Maximal Breadth of Skull', fontsize = 15)
plt.xlabel('Year', fontsize = 10)
plt.ylabel('Maximal Breadth of Skull', fontsize = 10)
plt.show()
```



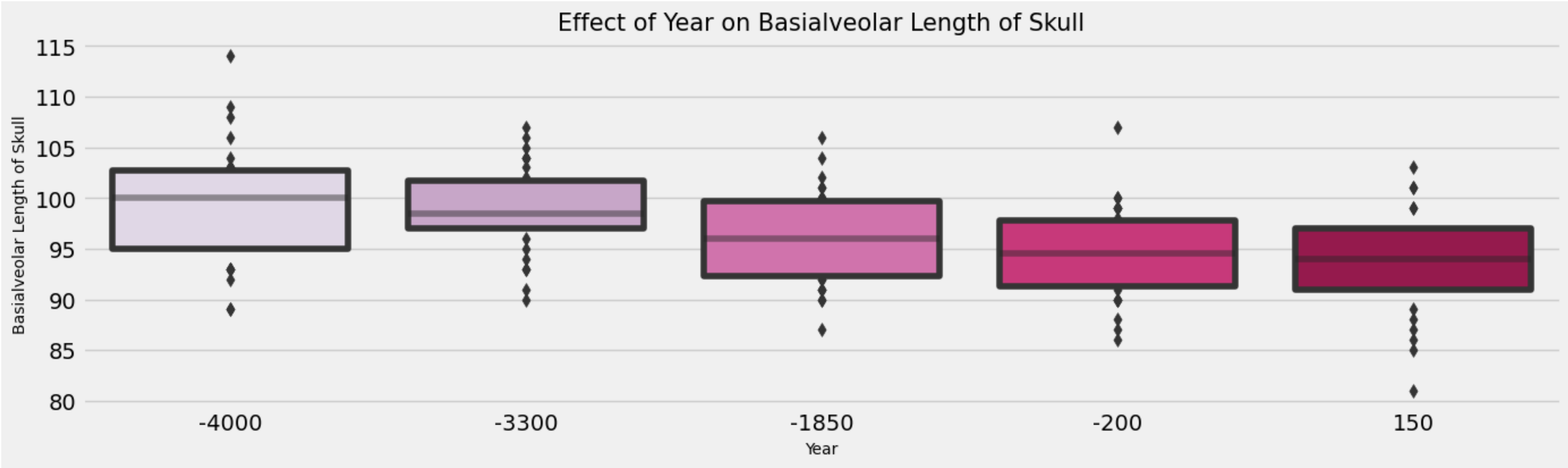
From the Figure above we can conclude that the Maximal Breadth of Skull is getting higher as the years goes on.

In [12]:
```
# Effect of Year on Basibregmatic Height of Skull

plt.rcParams['figure.figsize'] = (15,4)
sns.boxenplot(x = df['Year'].values,y= df['Basibregmatic Height of Skull'].values, palette = 'PuRd')
plt.title('Effect of Year on Basibregmatic Height of Skull', fontsize = 15)
plt.xlabel('Year', fontsize = 10)
plt.ylabel('Basibregmatic Height of Skull', fontsize = 10)
plt.show()
```



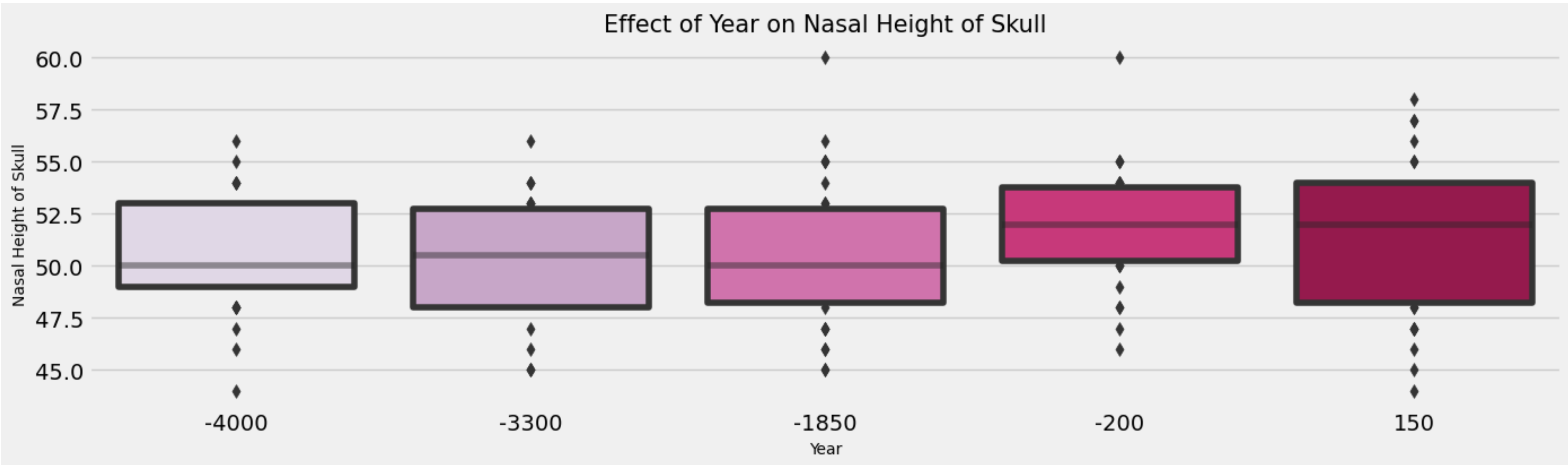
From the Figure above we can conclude that the Basibregmatic Height of Skull is almost the same during all time.

In [13]:
```
# Effect of Year on Basialveolar Length of Skull

plt.rcParams['figure.figsize'] = (15,4)
sns.boxplot(x = df['Year'].values,y= df['Basialveolar Length of Skull'].values, palette = 'PuRd')
plt.title('Effect of Year on Basialveolar Length of Skull', fontsize = 15)
plt.xlabel('Year', fontsize = 10)
plt.ylabel('Basialveolar Length of Skull', fontsize = 10)
plt.show()
```



From the Figure above we can conclude that the Basialveolar Length of Skull is getting lower as the years goes on.

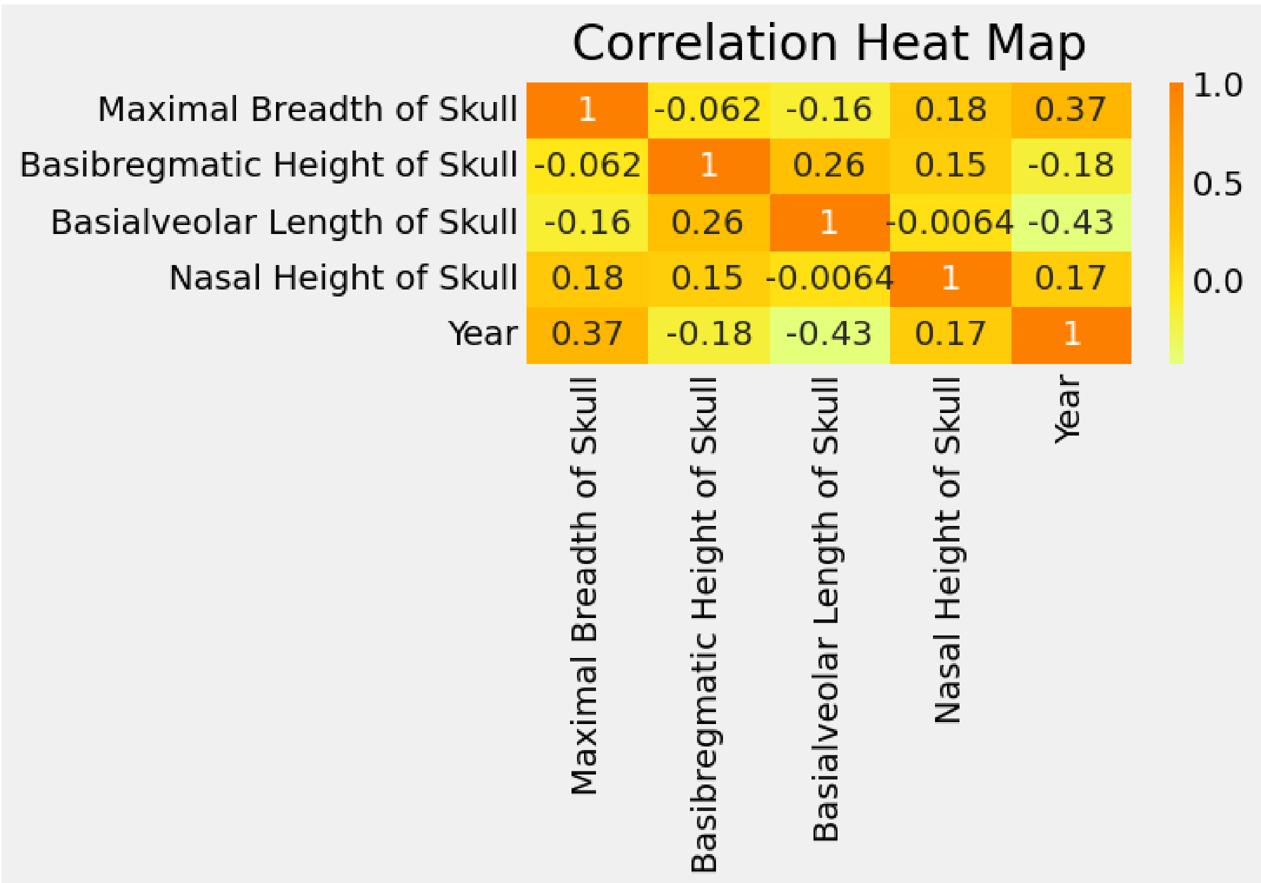
In [14]:
```
# Effect of Year on Nasal Height of Skull

plt.rcParams['figure.figsize'] = (15,4)
sns.boxplot(x = df['Year'].values,y= df['Nasal Height of Skull'].values, palette = 'PuRd')
plt.title('Effect of Year on Nasal Height of Skull', fontsize = 15)
plt.xlabel('Year', fontsize = 10)
plt.ylabel('Nasal Height of Skull', fontsize = 10)
plt.show()
```



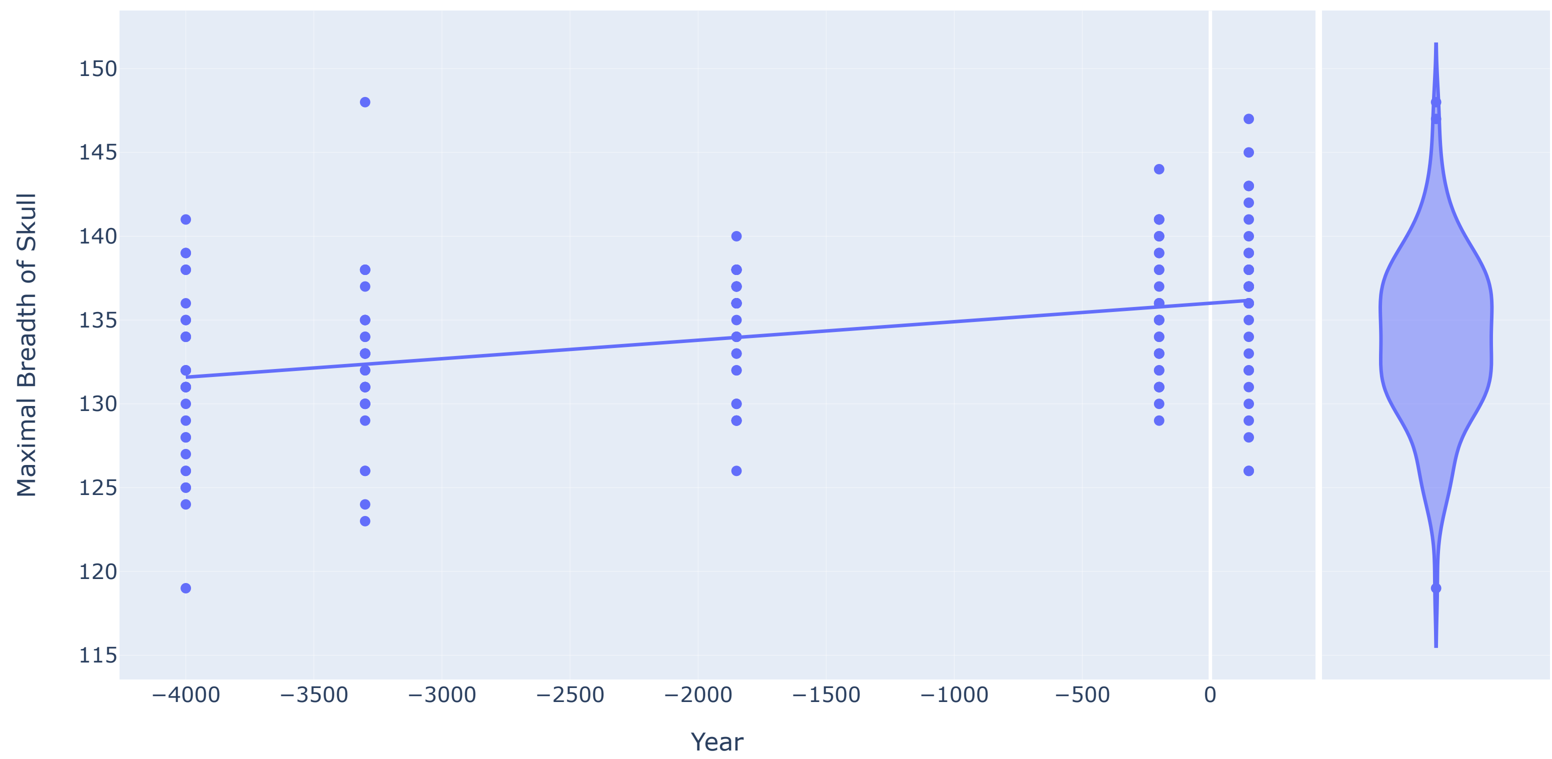
From the Figure above we can conclude that the Nasal Height of Skull is getting higher as the years goes on.

In [15]:
```
#
plt.figure(figsize=(5,2))
plt.title("Correlation Heat Map", y = 1.03)
sns.heatmap(df.corr(), cmap='Wistia',annot=True)
```

Out[15]: <AxesSubplot:title={'center':'Correlation Heat Map'}>

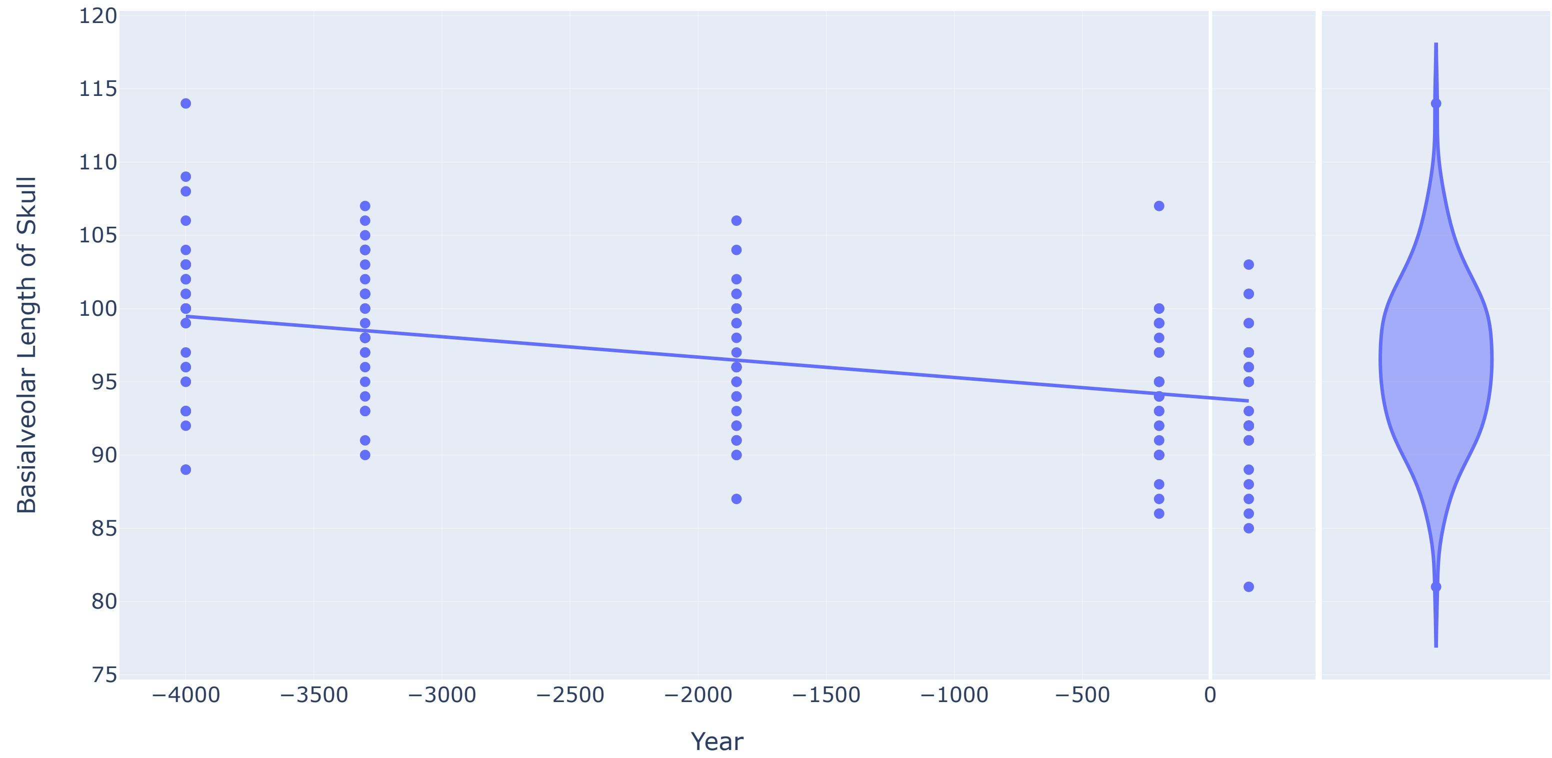


In [16]:
```
# lets understand the impact of Year on Maximal Breadth of Skull
px.scatter(df, y='Maximal Breadth of Skull',x='Year',
           marginal_y = 'violin',
           trendline = 'ols')
```



We can see that without the outlier in the year -3300, there is a clear improvement during the year when talking about Maximal Breadth of Skull.

In [17]:
```
# lets understand the impact of Year on Basialveolar Length of Skull
px.scatter(df, y='Basialveolar Length of Skull',x='Year',
           marginal_y = 'violin',
           trendline = 'ols')
```



We can conclude that during the years Basialveolar Length of Skull is going down.