In [115]:

```
import pandas as pd
import matplotlib as mlp
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
pd.set_option("display.precision", 2)

mlp.__version__
```

Out[115]:

```
'3.5.3'
```

Dataset Setup

Here I will start by extracting data into the program and create a dataframe to house its dataset. Here I also in the process to look at the data and possibly clean the data before I use it.

In [116]:

```
census_dataset = 'census2000.csv'
#read the given dataset extracted from the census2000 csv file
df_census = pd.read_csv(census_dataset)

df_census.head()
```

Out[116]:

	Sex	Year	Age	People
0	1	1900	0	4619544
1	1	2000	0	9735380
2	1	1900	5	4465783
3	1	2000	5	10552146
4	1	1900	10	4057669

In [ ]:

```
```

Here we see the dataframe and visualize the first 5 data rows in the data and check if it runs.

In [117]:

```
df_census_1900 = df_census[df_census['Year']== 1900]
#here I split the data between 1900 and 2000
df_census_2000 = df_census[df_census['Year']== 2000]
#visualize the dataframe
df_census_2000.head()
```

Out[117]:

	Sex	Year	Age	People
1	1	2000	0	9735380
3	1	2000	5	10552146
5	1	2000	10	10563233
7	1	2000	15	10237419
9	1	2000	20	9731315

In [118]:

```
df_census_1900.head(20)
```

Out[118]:

	Sex	Year	Age	People
0	1	1900	0	4619544
2	1	1900	5	4465783
4	1	1900	10	4057669
6	1	1900	15	3774846
8	1	1900	20	3694038
10	1	1900	25	3389280
12	1	1900	30	2918964
14	1	1900	35	2633883
16	1	1900	40	2261070
18	1	1900	45	1868413
20	1	1900	50	1571038
22	1	1900	55	1161908
24	1	1900	60	916571
26	1	1900	65	672663
28	1	1900	70	454747
30	1	1900	75	268211
32	1	1900	80	127435
34	1	1900	85	44008
36	1	1900	90	15164
38	2	1900	0	4589196

Here I am just making sure that it included both male and female data with its given year. The method I am doing here is to separate these two century type years from each other and then split them again between gender. Also I am currently playing around with data and see what works before using some of them to create my question and visualization. I am recording my results and process of transforming or cleaning data to my liking. It is good to let the audience know what I did with my data and the process to get to my result.

In [119]:

```
#here I split male and female between the years
df_1900_1 = df_census_1900[df_census_1900['Sex']== 1]

df_1900_2 = df_census_1900[df_census_1900['Sex']== 2]

df_2000_1 = df_census_2000[df_census_2000['Sex']== 1]

df_2000_2 = df_census_2000[df_census_2000['Sex']== 2]
```

The split is to give me more options of change if I need to use it. This can also be used for brain storming purposes and to view what the dataframe look like individually for each age or gender. In practice that means I am not using these to create my graph but to visualize before changing any data. And then here I check one of the dataframes to make sure the years and the sex are the same and cleaned in preparation for my visualization.

In [120]:

```
df_1900_1.head(20)
```

Out[120]:

	Sex	Year	Age	People
0	1	1900	0	4619544
2	1	1900	5	4465783
4	1	1900	10	4057669
6	1	1900	15	3774846
8	1	1900	20	3694038
10	1	1900	25	3389280
12	1	1900	30	2918964
14	1	1900	35	2633883
16	1	1900	40	2261070
18	1	1900	45	1868413
20	1	1900	50	1571038
22	1	1900	55	1161908
24	1	1900	60	916571
26	1	1900	65	672663
28	1	1900	70	454747
30	1	1900	75	268211
32	1	1900	80	127435
34	1	1900	85	44008
36	1	1900	90	15164

In [ ]:

In [ ]:

Data Question

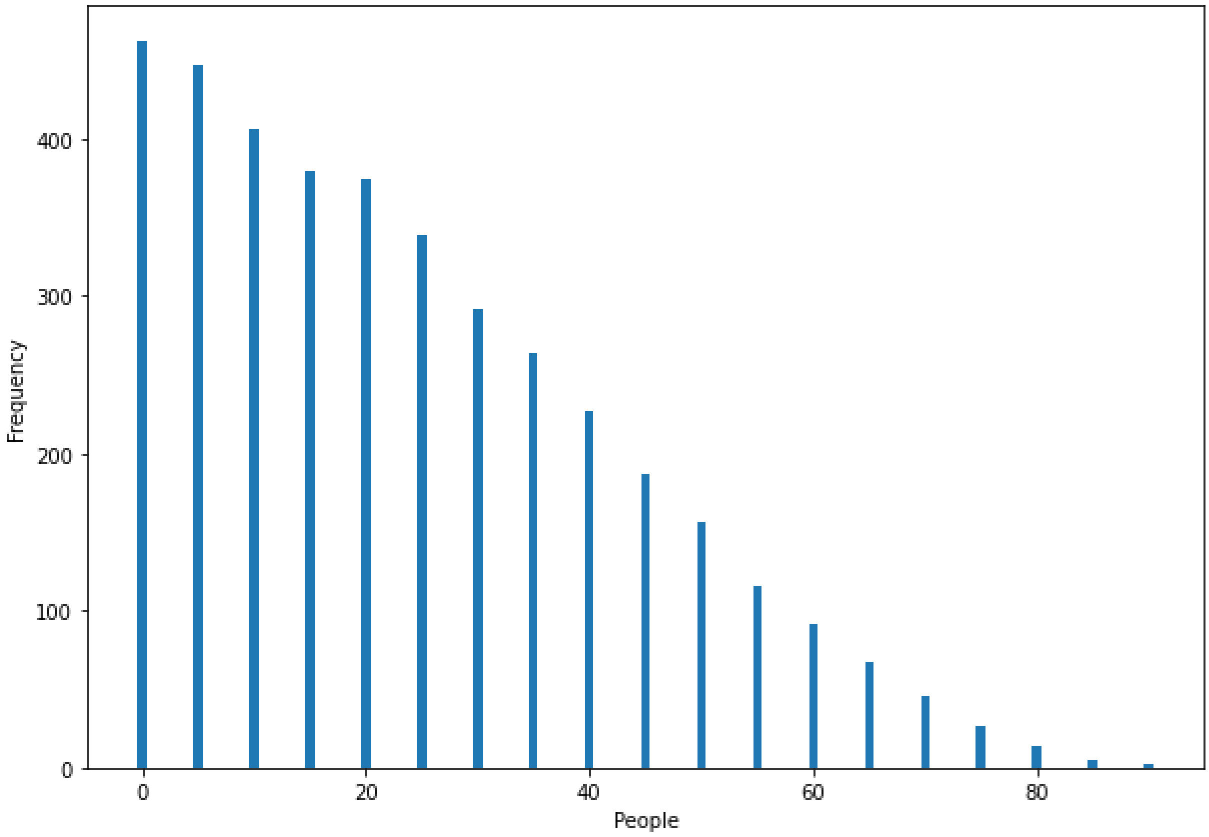
The question I will be posing in this assignment is "Which of the following sexs in its affiliated year (1900 and 2000) has the highest Frequency?" I will be figuring out which gender had the highest frequency in the 1900 and which gender is highest in the 2000.

Here I start off by creating a histogram to visualize one of the years to get a better view of what it looks like.

I will first visualize the histogram of the 1900 for both male and female.

In [170]:

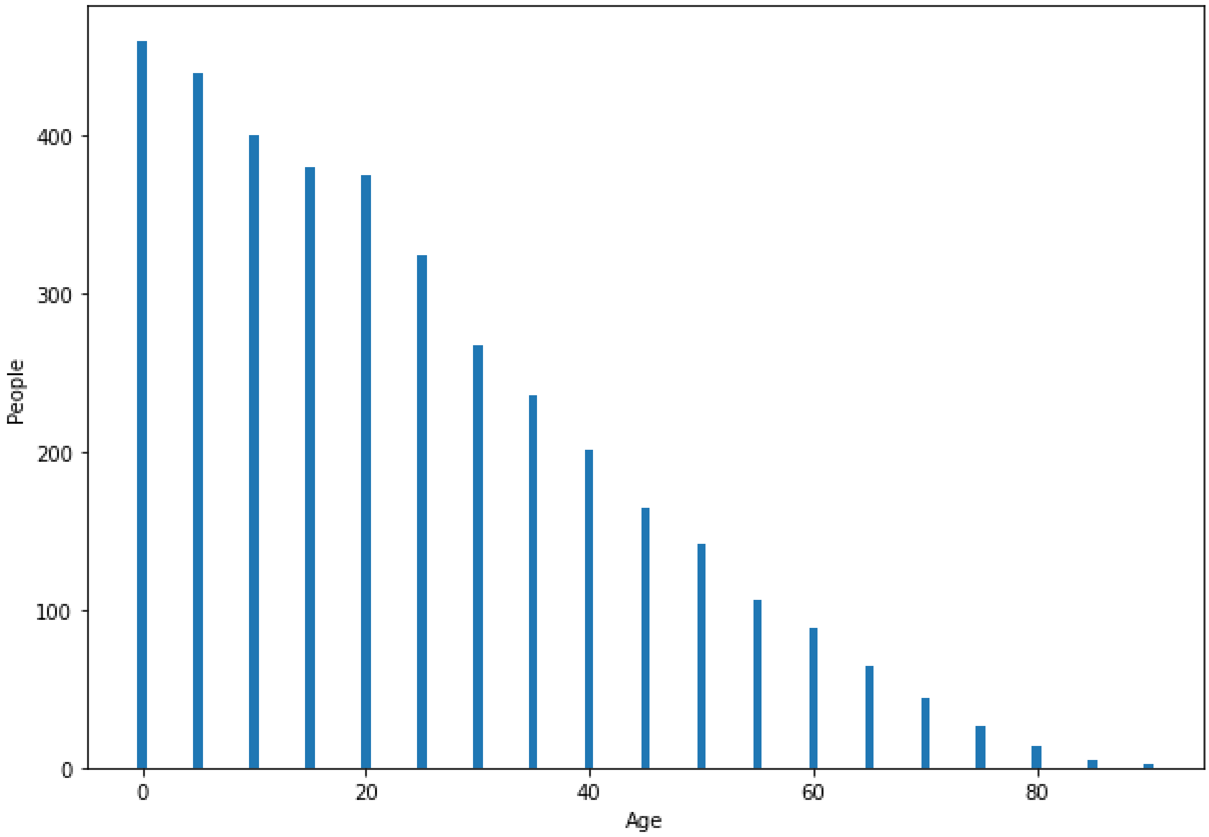
```
#n_bins = 100 #number of bins
fig, ax = plt.subplots(figsize=(10, 7))
#ax.hist(df_census_1900['People'], bins=n_bins, edgecolor='k')

ax.bar(df_census_1900['Age'], df_census_1900['People']/10**4) #People is divided

ax.set_xlabel('Age')
ax.set_ylabel('People')

plt.show() # show the plot
```



Upon visualizing the graph I realize that the graph plots the first half or the male version of the 1900 census as it is in the order of the data frame. So, I will be visualizing the female one next.

In [183]:

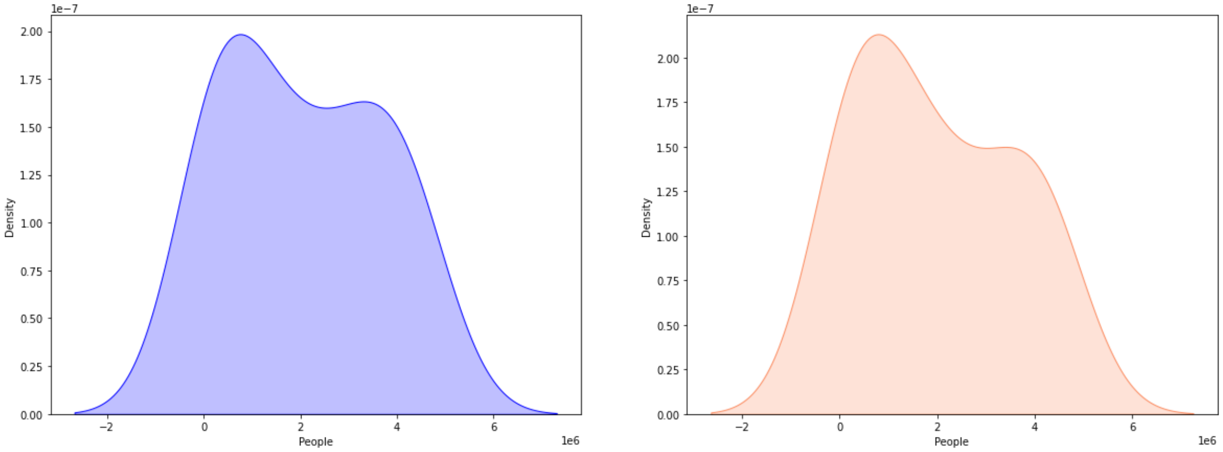
```
fig, ax = plt.subplots(figsize=(10, 7))

ax.bar(df_1900_2['Age'], df_1900_2['People']/10**4) #People is divided all by 10

ax.set_xlabel('Age')
ax.set_ylabel('People')

plt.show() # show the plot
```



Upon visualizing the graph, now that I see that the graphs look similar. So I will create a Density/Single Distribution Plot to better my visualization. Forming a density plot will better my understanding of the two Sexes and separate them apart to better look at how they are affected.

In [180]:

```
span = df_census_1900['People'].max() - df_census_1900['People'].min()

bin_width = 5
n_bins = int(span/bin_width)
fig, (ax1, ax2) = plt.subplots(1,2, figsize = (20,7))
#sns.kdeplot(df_census_1900.loc[df_census_1900['Sex'] == 1 ,'People'], fill=True,
#sns.kdeplot(df_census_1900.loc[df_census_1900['Sex'] == 2 ,'People'], fill=True,
sns.kdeplot(df_census_1900.loc[df_census_1900['Sex'] == 1 ,'People'], fill=True,
sns.kdeplot(df_census_1900.loc[df_census_1900['Sex'] == 2 ,'People'], fill=True,


plt.show()
```

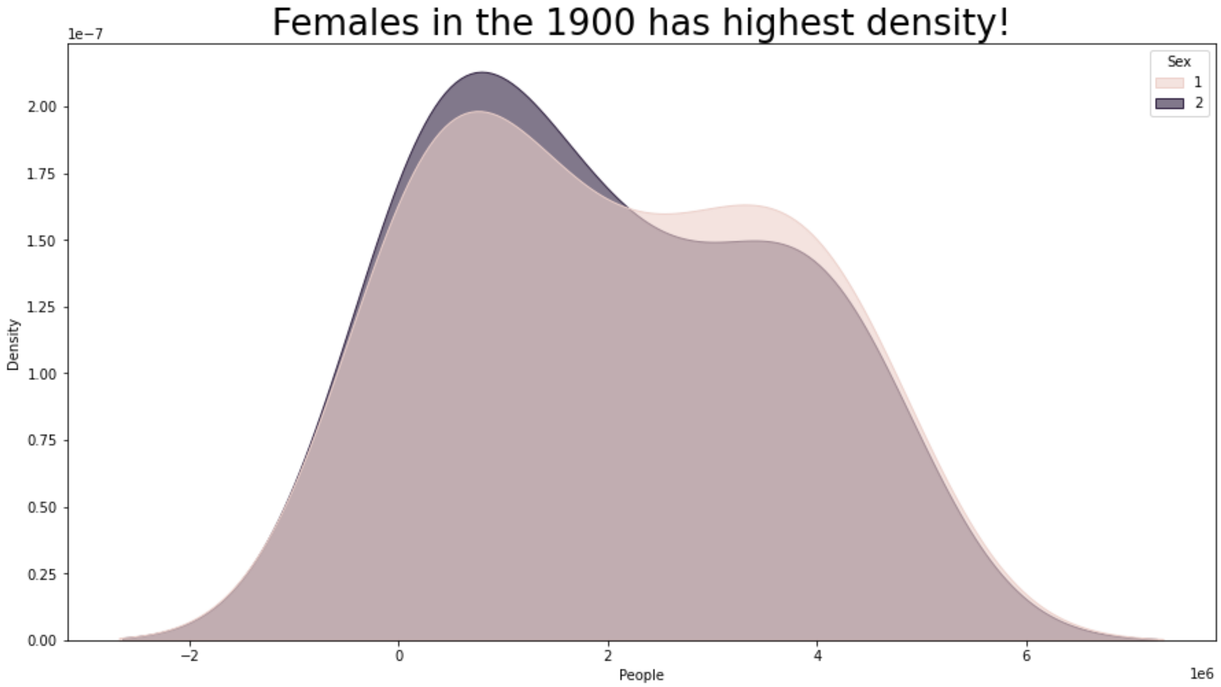


I have Put the Sexes seperately as this will give me an overview of what Density they have. Upon inspection it appears that the female is dominant in this year but the graph does not convey this as the diameter and parameter are made to fit to the graph. It doesn't represent its actual density and makes it hard to know which is higher than the other.

Next I will be create a multiple distribution featuring these two graphs combined and overlapped to each other.

In [235]:

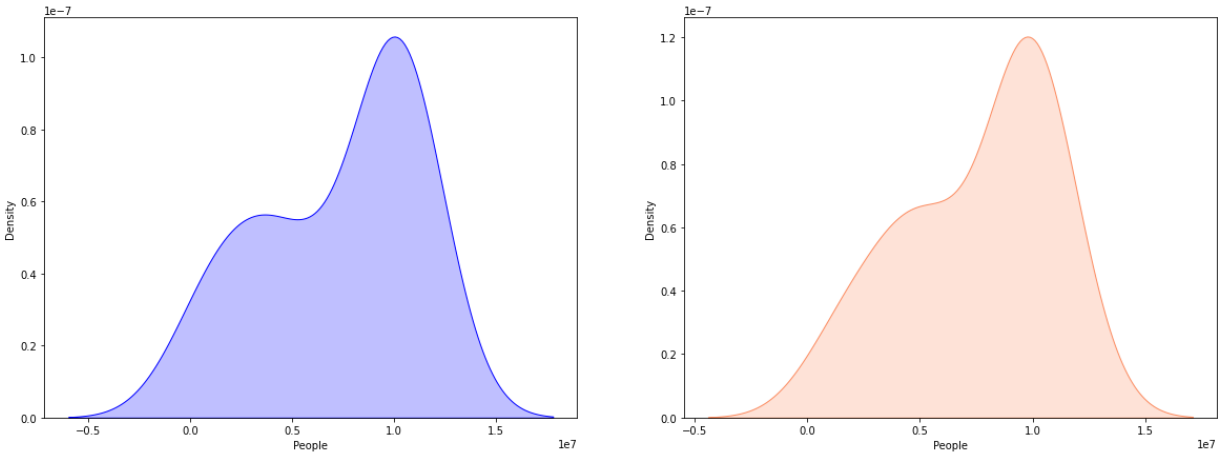
```
span = df_census_1900['People'].max() - df_census_1900['People'].min()
bin_width = 5 #five year
n_bins = int(span/bin_width)
fig, ax = plt.subplots(figsize=(15,8))
sns.kdeplot(data=df_census_1900, x='People', hue='Sex', fill=True, common_norm=Fa

ax.set_title('Females in the 1900 has highest density!', fontsize=26, loc='center

plt.show() # show the plot
```



Now looking at the graph it is made possible that the visualization shows that woman or females had the highest densities compared to men. Using overlapping design helps to visualize which gender had the highest density or frequency during that year. This tests that females reached the highest density in 1900.

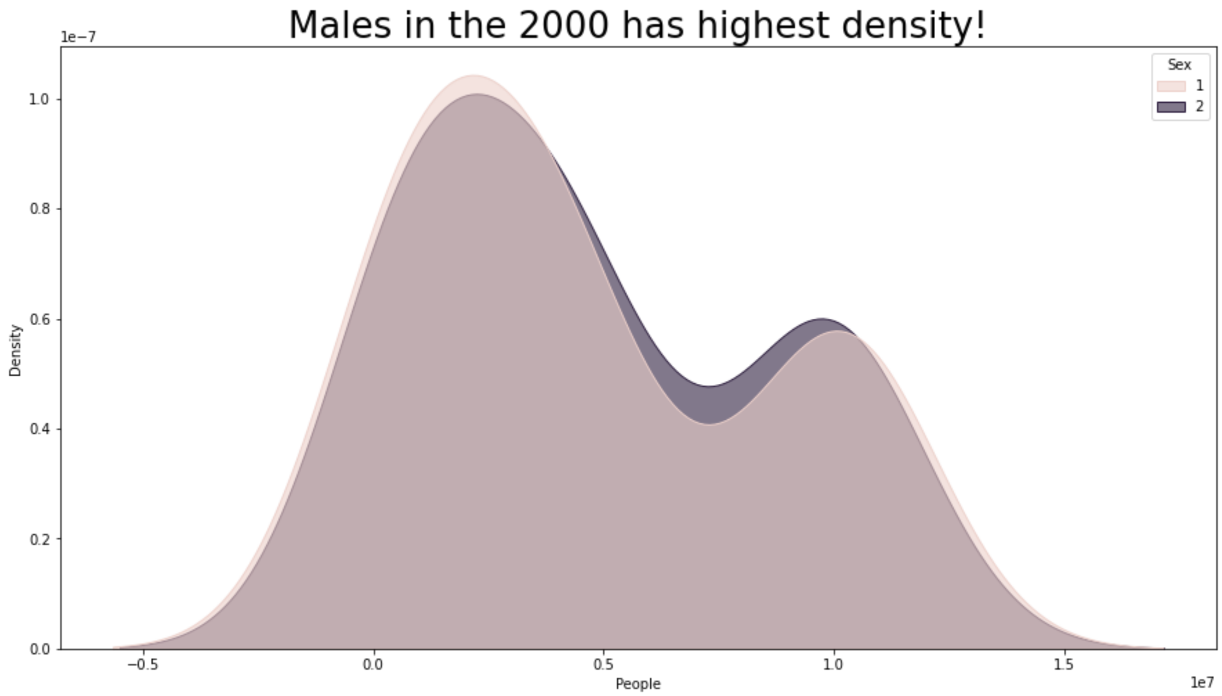
I will do the same to the year 2000 and see who has the highest among its sex category.

In [221]:
```
span = df_census_2000['People'].max() - df_census_2000['People'].min()

bin_width = 5
n_bins = int(span/bin_width)
fig, (ax1, ax2) = plt.subplots(1,2, figsize = (20,7))

sns.kdeplot(df_census_2000.loc[df_census_2000['Sex'] == 1 ,'People'], fill=True,
sns.kdeplot(df_census_2000.loc[df_census_2000['Sex'] == 2 ,'People'], fill=True,


plt.show()
```



In [238]:

```
span = df_census_2000['People'].max() - df_census_2000['People'].min()
bin_width = 5 #five year
n_bins = int(span/bin_width)
fig, ax = plt.subplots(figsize=(15,8))
sns.kdeplot(data=df_census, x='People', hue='Sex',fill=True, common_norm=False, a
ax.set_title('Males in the 2000 has highest density!', fontsize=26, loc='center')

plt.show() # show the plot
```



In 2000, Males had the highest density and overtakes a bit of females during that time.

Add alternative text

"A Density Plot graph displaying density amounts of the set gender for that given year. Males and Females overall accumulation of People that are born and populated during that year. Many of the People in their given sex category has similar densities. But, I look for highest as to look into which sex reached highest density in a given year"

Employ a takeaway title

Here, I will display 2 different graphs from two different years as I want to display which had the highest density in that year. The title "Females in the 1900 has highest density!" and "Males in the 2000 has highest density!".

Label Data Directly:

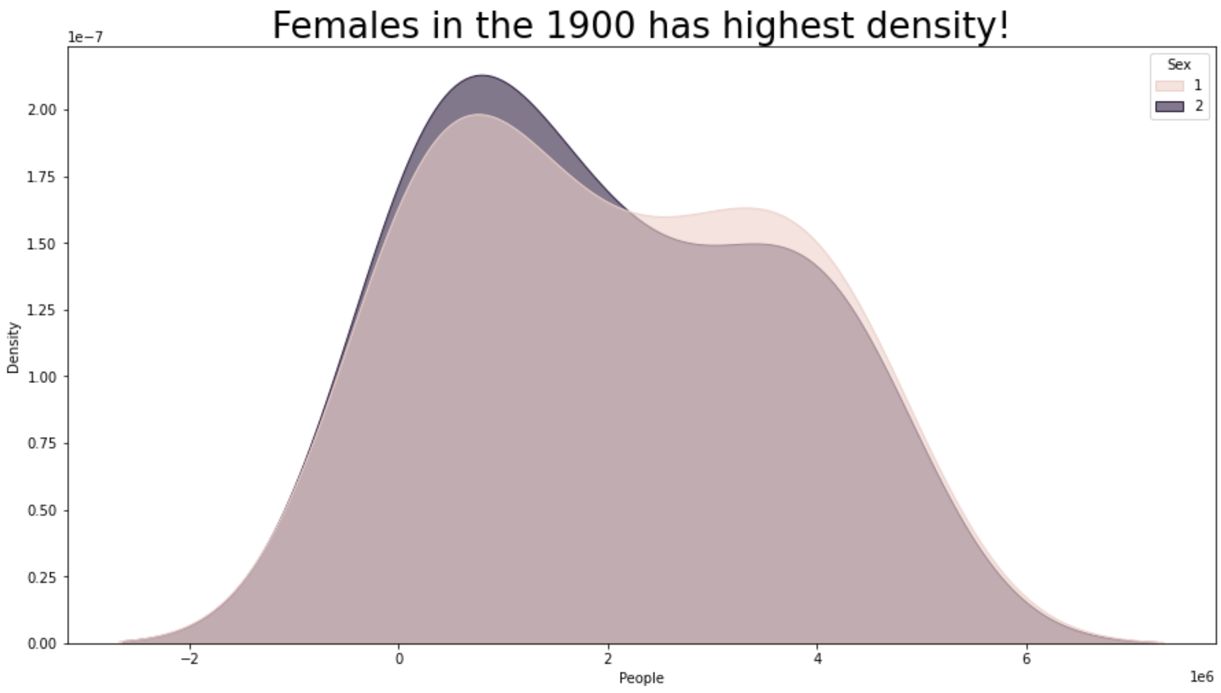
Currently there is a legend at the corner of the graph that displays the colors given a shade color and a darker color. The given hue makes it that it cannot be easily changed. By default, "People" label is embedded in the graph indicates in a quarterly fashion where it starts at 0 where its highest and then 2 where that indicates 100 years of age. This also has a x-y relationship from People with Age as it shows correlation between People and its affiliated density.

Use sufficient contrast¶

The graph consist of 2 colors only and has a shade of dark and light. This comes from the Hue coding and is defaulted that way where it can't be changed as well.

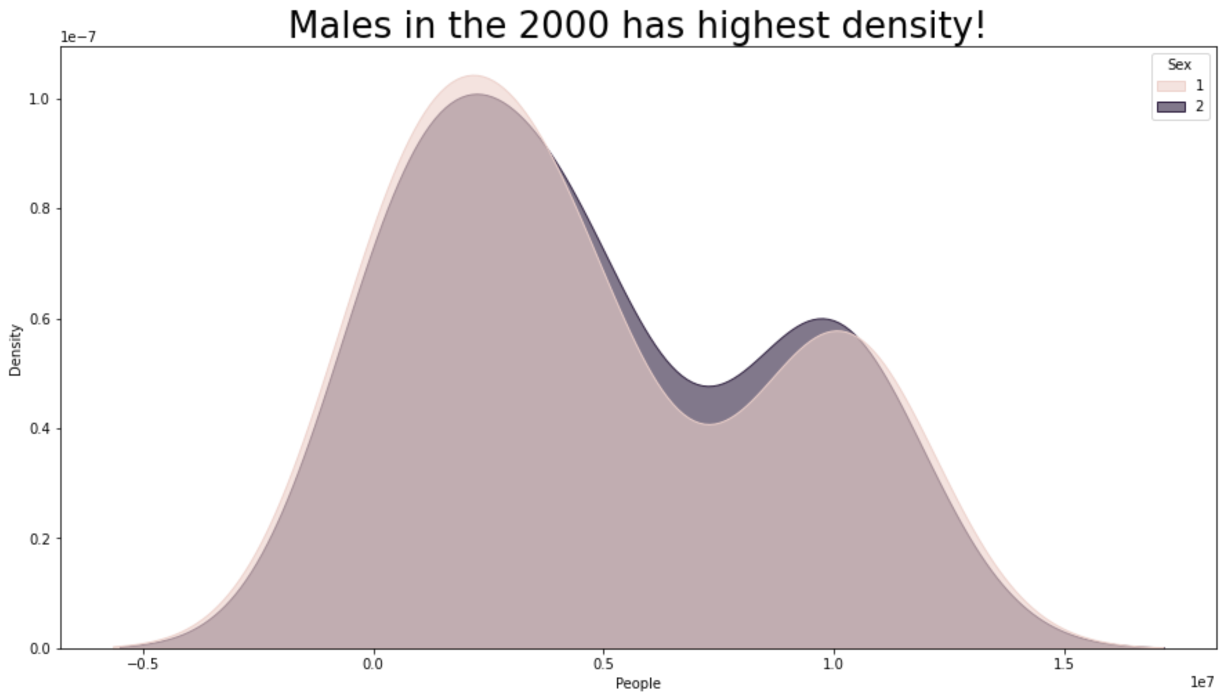
Final Improved for Accessibility Plot

```
In [239]:
span = df_census_1900['People'].max() - df_census_1900['People'].min()
bin_width = 5 #five year
n_bins = int(span/bin_width)
fig, ax = plt.subplots(figsize=(15,8))
sns.kdeplot(data=df_census_1900, x='People', hue='Sex', fill=True, common_norm=Fa

ax.set_title('Females in the 1900 has highest density!', fontsize=26, loc='center

plt.show() # show the plot
```



In [240]:

```
span = df_census_2000['People'].max() - df_census_2000['People'].min()
bin_width = 5 #five year
n_bins = int(span/bin_width)
fig, ax = plt.subplots(figsize=(15,8))
sns.kdeplot(data=df_census, x='People', hue='Sex',fill=True, common_norm=False, a
ax.set_title('Males in the 2000 has highest density!', fontsize=26, loc='center')

plt.show() # show the plot
```



The Plot shows the finished product where there are two separate but, engaging titles with a forward and progressive statement that engages the audience. There are paths where I had to switch from one graph to the other and choose the one that will fit my graph and story best.

From the beginning I screened the data and look at individual lines to see what the code and data is all about. Knowing that it came with 2 genders I proceeded to use the dataframe to seperate them and view them in their designated year and order. I looked at the data and knew that the graph follows a x-y relationship with Age and People. I utilized People as my X-axis to reflect that and not use Age. Reason I did this is that the People and Age category go hand in hand and reflect the same outcome. When looking at the finished product, it starts at 0 and ends at 2. This is similar to a quarter graph where from 0-25 age wise it is 0-0.5 on the "People" x-axis graph. There is another graph that has it up to 10 and uses the same concept but a slightly different scale. I did a different approach from having a y-axis for "People" and x-axis for "Age" and went for a different method to look into highest number. Knowing this I proceeded with brainstoming technique and laid every different dataframe in different years and genders. This was my template to start on the visualization process.

I went to create my question a bit later in the process, I first started with a histogram graph using the manual way of creating a graph. X-axis for Age and Y-axis for People. Looking at this the graph came our rather monotonic and thin. I proceeded to try another graph of the same code but, using different sex category. They still looked the same. I resulted from this that histograms is not the way to go and went to do Density plots.

When producing the plots I first went and use the 1900 dataset to test and visualize the graph to see if it was compatible and easier to read. It was a lot easier to read but the parameters and angles of the graph still looked the same but the y and x axis numbers were differet. The graph defaulted to fitting the graph and distorted the density plot for uniformity.

I then resulted to doing a combined multi-distribution plot to overlap and connect the two graphs to make the visualization display both graphs in their original state. And it worked. Although I have to note that the code and design I used may not accept many color changes as it only recognizes hue default colors. The colors could have changed to a more contrast color but with the male and female being numerical it can't be changed unless the graph drops the category and then replaces it with a string that it can read and change the "hue" color. This graph has a slight fixed structure to it which gives it convience to graph but, not so much on changing. I have found that this answered for the two century apart years, which of the two had the highest density. I then did the same with the year 2000 dataframe and visualized it.

```
In [ ]:
```

```
In [ ]:
```