

Kernel Density Estimates (KDEs) and generative models in Python

Interactive Lecture
Smith College

Dr Meridith Joyce

Marie Curie Widening Fellow: MATISSE
CSFK Konkoly Observatory, Budapest

MESA Developers



@MeridithJoyceGR

www.meridithjoyce.com

github.com/mjoyceGR

Big Picture:

Given an observed distribution of stellar ages in the center of our Galaxy, we would like to predict the distribution of stellar ages in other Galaxies

Big Picture:

Given an observed distribution of stellar ages in the center of our Galaxy, we would like to predict the distribution of stellar ages in other Galaxies

In this interactive lecture, we will apply a *kernel density estimate* to a measured stellar age distribution and use this to make synthetic data sets from a *generative model*

Step 0: Bookkeeping -- Google Drive

(1) Scan this QR code, save the link, then open it on the device you will use to code

You should see a Google Drive folder called Smith_KDE_Exercises

(2) download and save

- a) **KDE_exercises.ipynb**
- b) **KDE_full_solutions.ipynb**
- c) **stellar_ages.dat**

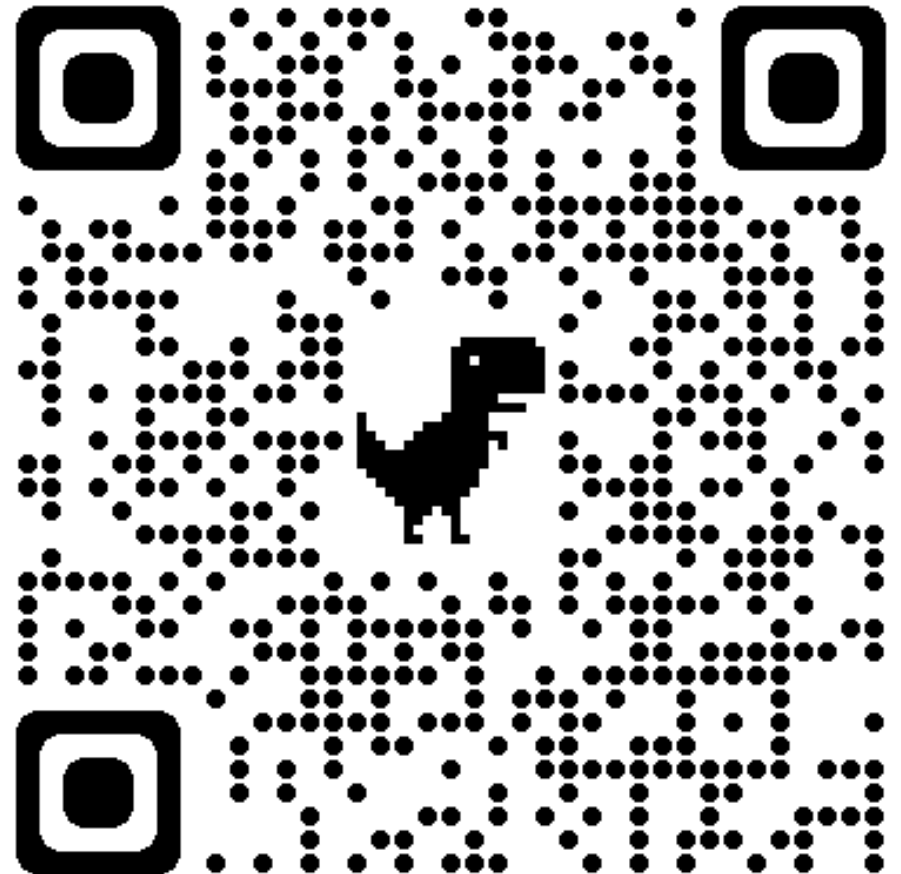
(3) Go to

<https://colab.research.google.com/>

(i) Under “File” in the top left, click “open notebook”

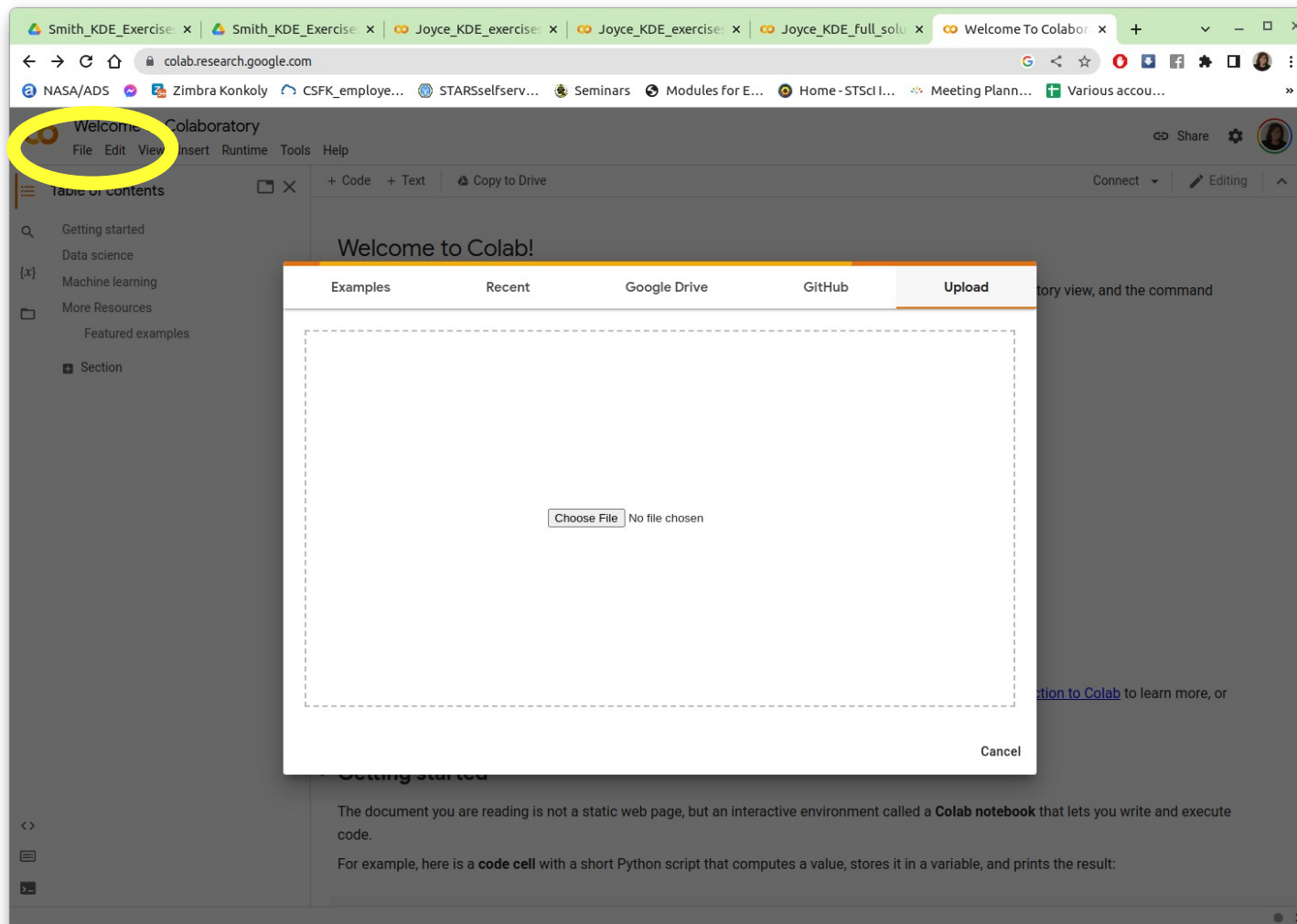
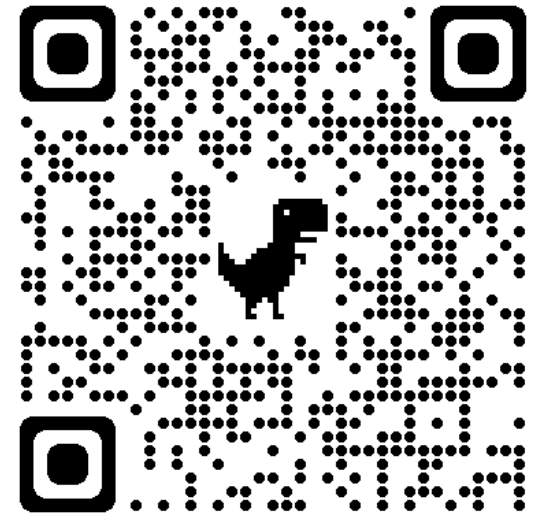
(ii) Click “Upload” on the far right of the menu that pops up

(iii) Upload files a, b, c



Step 0: Bookkeeping

<https://colab.research.google.com/>



Step 0: Bookkeeping -- Github/Jupyter

(1) go to https://github.com/mjoyceGR/KDE_exercise

(2) download and save

KDE_exercises.ipynb

KDE_full_solutions.ipynb

stellar_ages.dat

in the same folder on your computer

(3) Load the two ipynb files into your preferred Jupyter Notebook environment. You will want both accessible so you can consult the solutions if you get stuck

Smith_KDE_Exercis... x | Smith_KDE_Exercis... x | Joyce_KDE_exercis... x | Joyce_KDE_exercis... x | Joyce_KDE_full_solu... x | KDE_exercises.ipynb x

colab.research.google.com/drive/1hUIXDzVHWoSdCP1KLXP97f0Bgp56Z6ll

NASA/ADS | Zimbra Konkoly | CSFK_employe... | STARSSelfserv... | Seminars | Modules for E... | Home - STScI I... | Meeting Plann... | Various accou...

KDE_exercises.ipynb

File Edit View Insert Runtime Tools Help Last saved at 6:03 PM

Comment Share

+ Code + Text

Connect Editing

↑ ↓ ↶ ↷ ↻ ↺ ↻ ↺

Step 1: Load the modules we will need

```
#!/usr/bin/env python3
#####
#
# template by M Joyce
# for use with Smith College students
#
#####

## import the modules
import numpy as np
import matplotlib.pyplot as plt
import scipy
from scipy import stats
from scipy.stats import norm

print("modules imported")
```

Is everyone on this page?

Step 2: Define a function to make figures look nice

```
[ ] def set_fig(ax):
    ax.tick_params(axis = 'both',which='both', width=2)
    ax.tick_params(axis = 'both',which='major', length=12)
    ax.tick_params(axis = 'both',which='minor', length=8, color='black')
    ax.tick_params(axis='both', which='major', labelsz=24)
    ax.tick_params(axis='both', which='minor', labelsz=20)
    return

print("plot settings function defined")
```

Step 3: Load the data

Pause

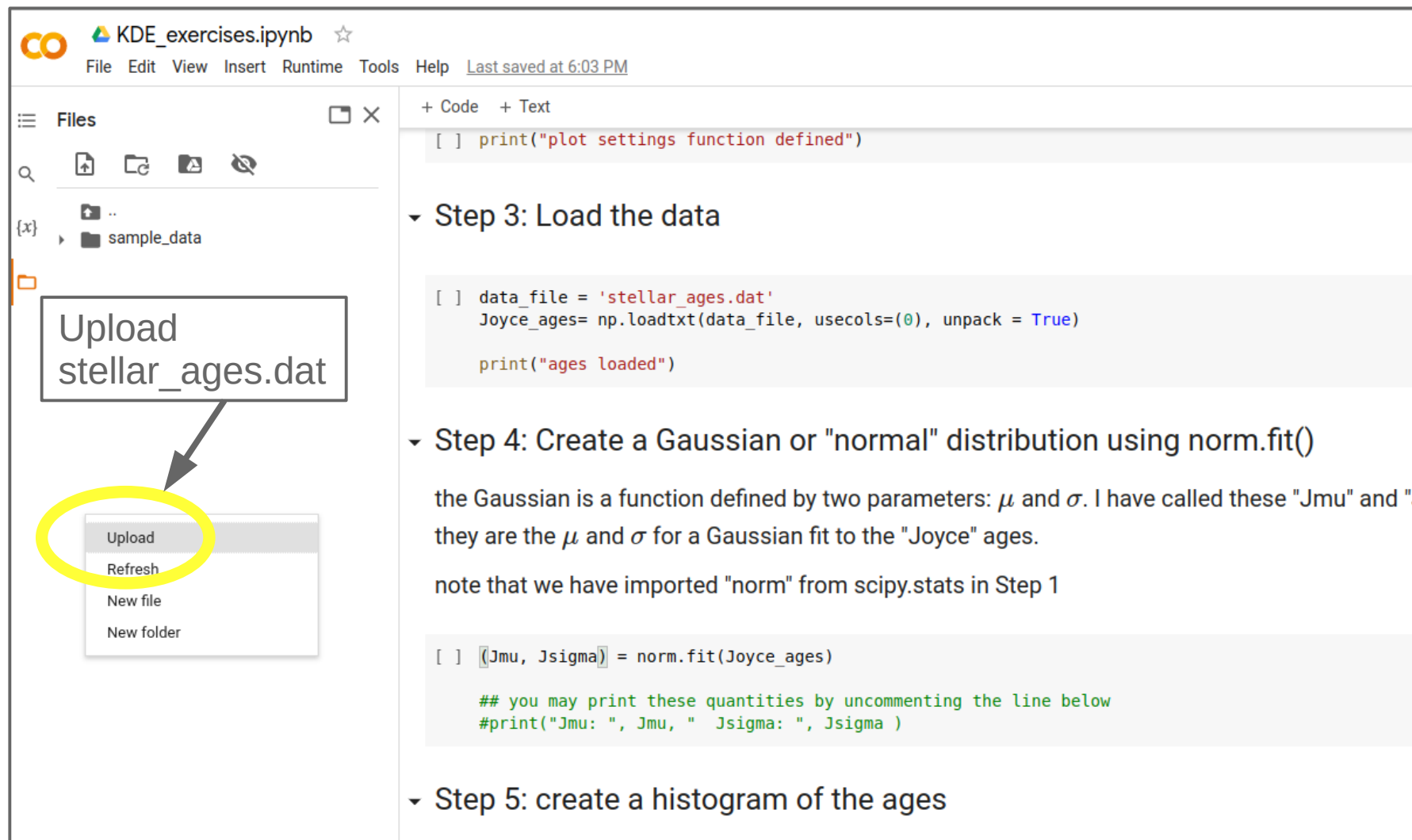
Step 1: Load the modules

Step 2: define a figure function `set_fig()`

You do not need to understand how this works, just that it makes the figures look nice

Step 3: Load the data

You must upload `stellar_ages.dat` in colab, *in* the notebook you have loaded (`KDE_exercises.ipynb`)



The screenshot shows a Google Colab notebook titled "KDE_exercises.ipynb". The left sidebar displays a file explorer with a folder named "sample_data". A text box with the text "Upload stellar_ages.dat" has an arrow pointing to the "Upload" button in the sidebar's context menu, which is highlighted with a yellow circle. The main notebook area contains the following code blocks:

```
[ ] print("plot settings function defined")
```

▼ Step 3: Load the data

```
[ ] data_file = 'stellar_ages.dat'
Joyce_ages= np.loadtxt(data_file, usecols=(0), unpack = True)

print("ages loaded")
```

▼ Step 4: Create a Gaussian or "normal" distribution using `norm.fit()`

the Gaussian is a function defined by two parameters: μ and σ . I have called these "Jmu" and "Jsigma" for a Gaussian fit to the "Joyce" ages.

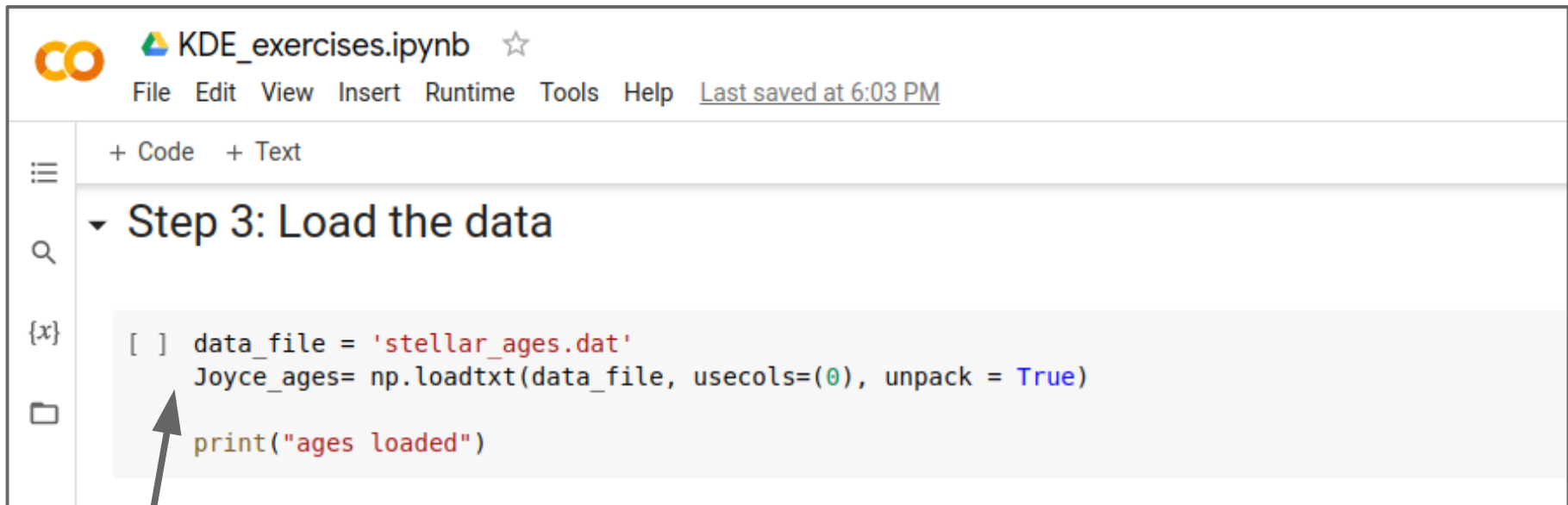
note that we have imported "norm" from `scipy.stats` in Step 1

```
[ ] (Jmu, Jsigma) = norm.fit(Joyce_ages)

## you may print these quantities by uncommenting the line below
#print("Jmu: ", Jmu, " Jsigma: ", Jsigma )
```

▼ Step 5: create a histogram of the ages

Step 3: (now) Load the data



The screenshot shows a Jupyter Notebook window titled 'KDE_exercises.ipynb'. The menu bar includes 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help', with a status bar indicating 'Last saved at 6:03 PM'. The left sidebar contains icons for a table of contents, search, variables, and files. The main area shows a code cell titled 'Step 3: Load the data' with the following Python code:

```
[ ] data_file = 'stellar_ages.dat'
    Joyce_ages= np.loadtxt(data_file, usecols=(0), unpack = True)

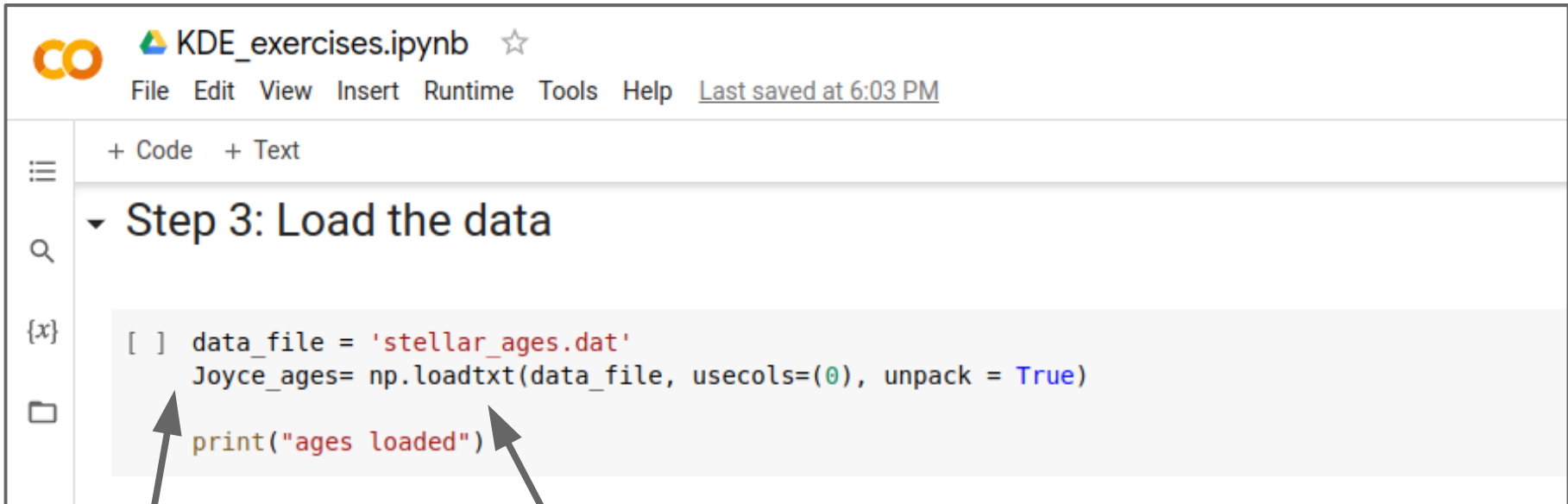
    print("ages loaded")
```

An arrow points from the first line of code to the explanatory text box below.

The data we are loading is a set of 91 stellar age determinations, measured in Gigayears (1 Gyr = 1 billion years = 10^9 years)

These are real data from my research, hence “Joyce ages”

Step 3: (now) Load the data



The screenshot shows a Jupyter Notebook window titled 'KDE_exercises.ipynb'. The menu bar includes 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help', with a status bar indicating 'Last saved at 6:03 PM'. The left sidebar contains icons for a table of contents, search, variables, and files. The main area shows a code cell titled 'Step 3: Load the data' with the following code:

```
[ ] data_file = 'stellar_ages.dat'
Joyce_ages= np.loadtxt(data_file, usecols=(0), unpack = True)

print("ages loaded")
```

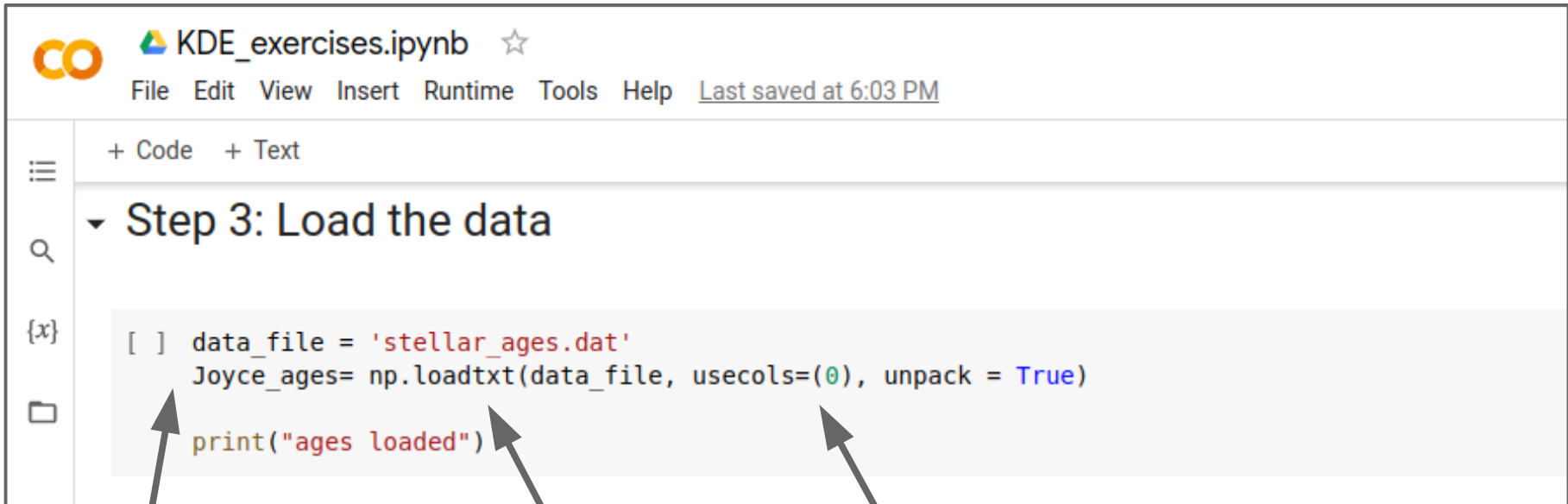
Two arrows point from the code to explanatory text boxes below. One arrow points from the file path 'stellar_ages.dat' to the first text box. The other arrow points from the 'np.loadtxt' function call to the second text box.

The data we are loading is a set of 91 stellar age determinations, measured in Gigayears (1 Gyr = 1 billion years = 10^9 years)

These are real data from my research, hence “Joyce ages”

We imported numpy as “np”
We are using a numpy function called “loadtxt” which automatically converts columns into np arrays

Step 3: (now) Load the data



The screenshot shows a Jupyter Notebook window titled 'KDE_exercises.ipynb'. The menu bar includes 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help', with a status bar indicating 'Last saved at 6:03 PM'. The left sidebar contains icons for a table of contents, search, and file explorer. The main area shows a code cell titled 'Step 3: Load the data' containing the following Python code:

```
[ ] data_file = 'stellar_ages.dat'
Joyce_ages= np.loadtxt(data_file, usecols=(0), unpack = True)

print("ages loaded")
```

Three arrows point from explanatory text boxes to specific parts of the code: one to the file name, one to the variable name, and one to the column index.

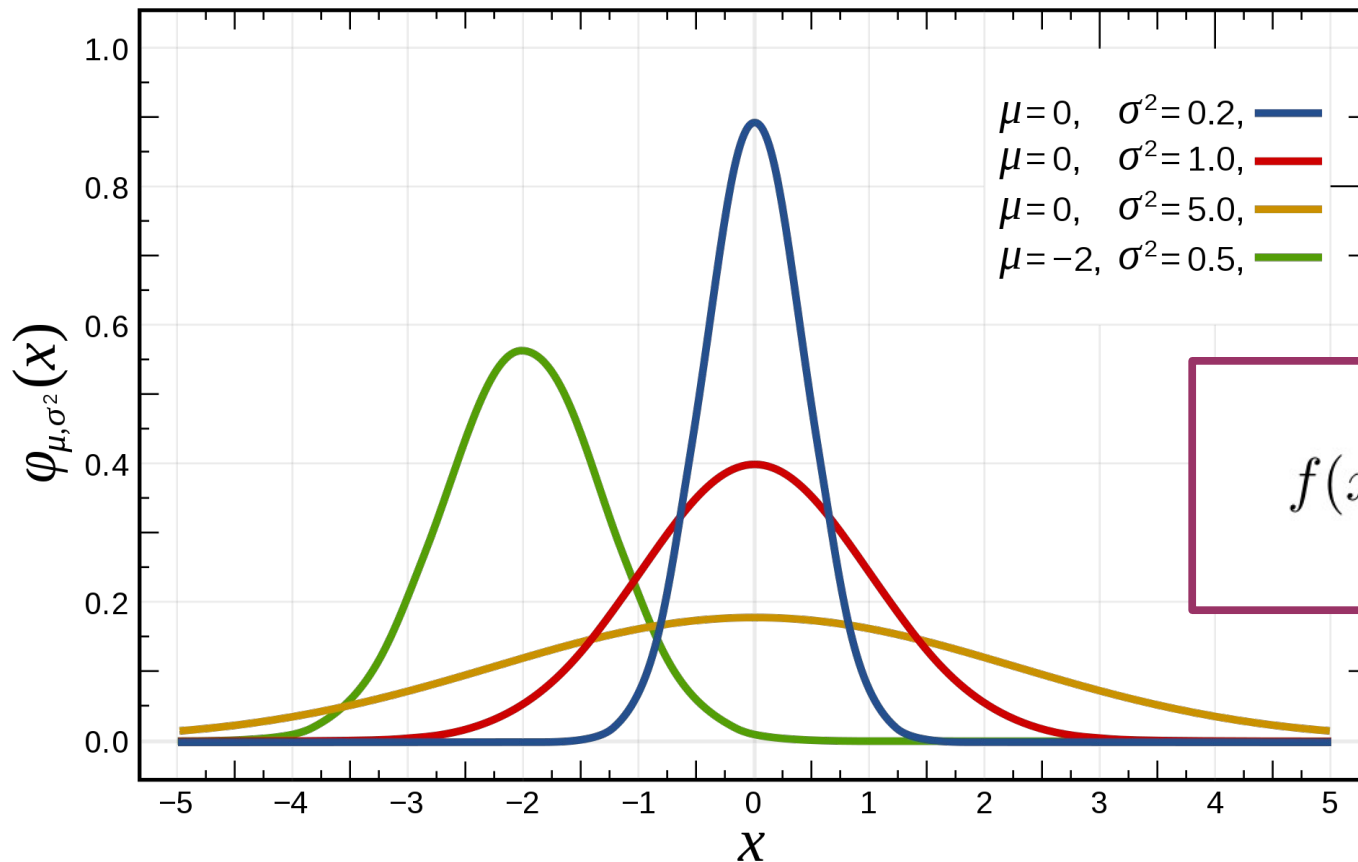
The data we are loading is a set of 91 stellar age determinations, measured in Gigayears (1 Gyr = 1 billion years = 10^9 years)

These are real data from my research, hence “Joyce ages”

Python indexes from zero, so the first column of stellar_ages.dat is “column 0”

We imported numpy as “np”
We are using a numpy function called “loadtxt” which automatically converts columns into np arrays

Step 4: Create a Gaussian



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- the shape of the curve is captured by **mu**, the “expected value,” or mean, and **sigma**, which is related to the width and represents one standard deviation (sigma² is the “variance,” as shown in the legend)
- the type of distribution everyone (in astronomy) assumes their data follow
- also called a “normal distribution”

Step 4: Create a Gaussian

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

▼ Step 4: Create a Gaussian or "normal" distribution using norm.fit()

the Gaussian is a function defined by two parameters: μ and σ . I have called these "Jmu" and "Jsigma" because they are the μ and σ for a Gaussian fit to the "Joyce" ages.

note that we have imported "norm" from scipy.stats in Step 1

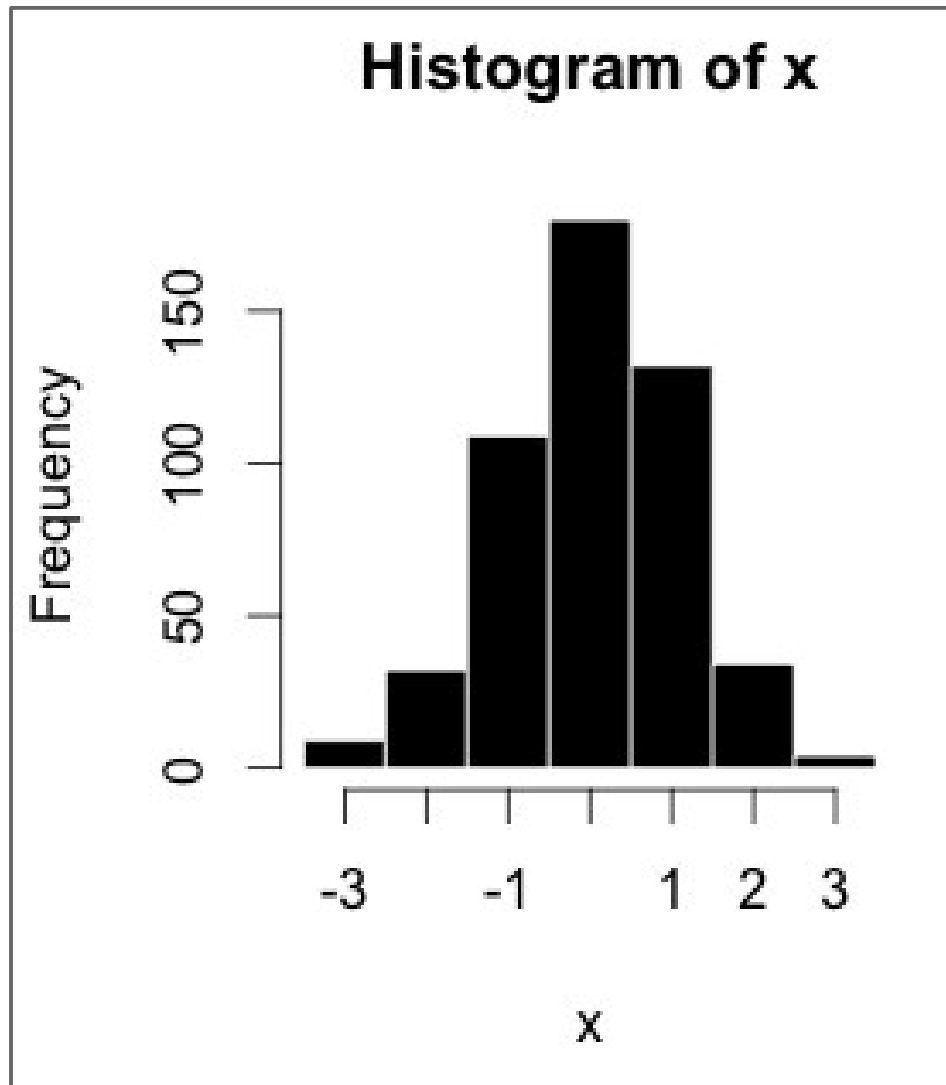
```
[ ] (Jmu, Jsigma) = norm.fit(Joyce_ages)

## you may print these quantities by uncommenting the line below
#print("Jmu: ", Jmu, " Jsigma: ", Jsigma )
```

- what this piece of code does is find the “best” values of mu and sigma for a fit of **$f(x)$** to the distribution formed by Joyce_ages

- it names these fit parameters *Jmu*, *Jsigma* and we will use them later to make a function

Step 5: Histogram and bins

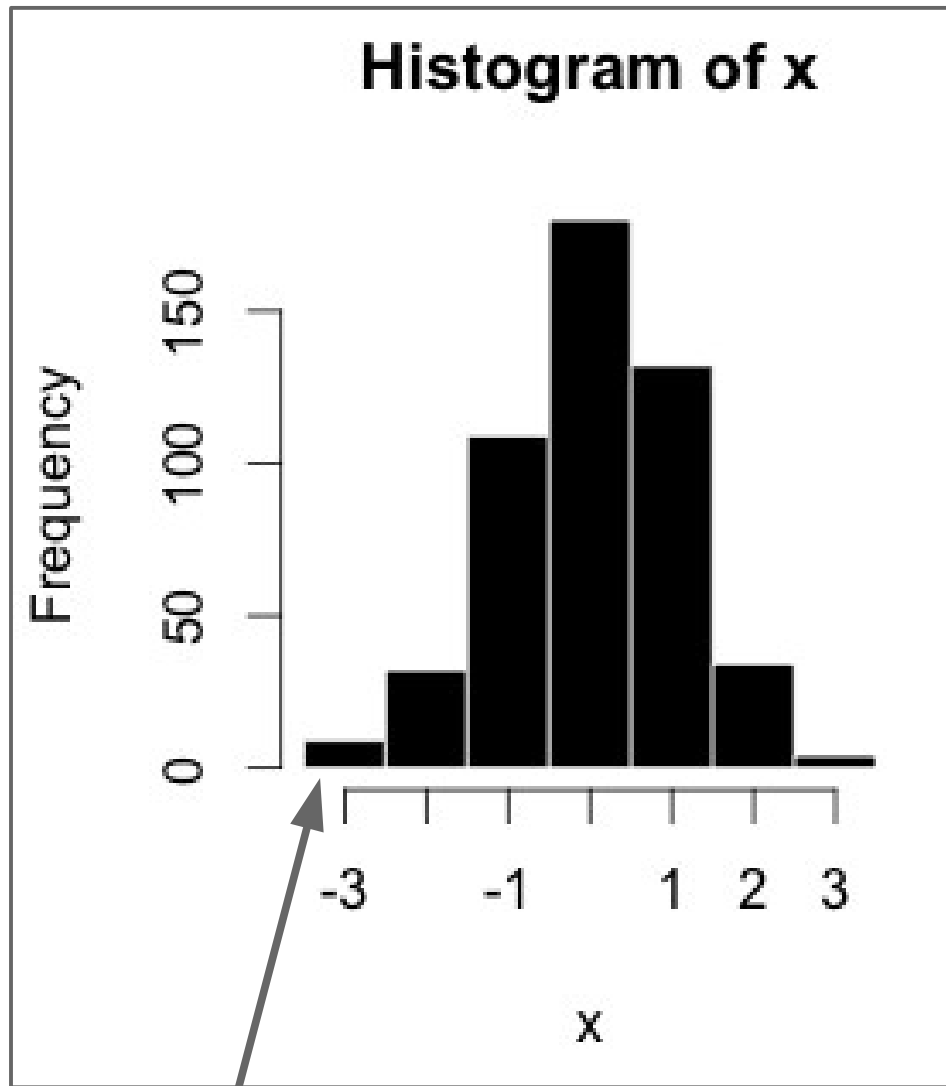


- on the x-axis, we have observations of some quantity
in our case, x = stellar ages

- the y-axis counts the number of occurrences of x

So, if out of 91 stars (x), we have 12 stars with an age of 10 Gyr, the y value for $x = 10$ will be 12.

Step 5: Histogram and bins



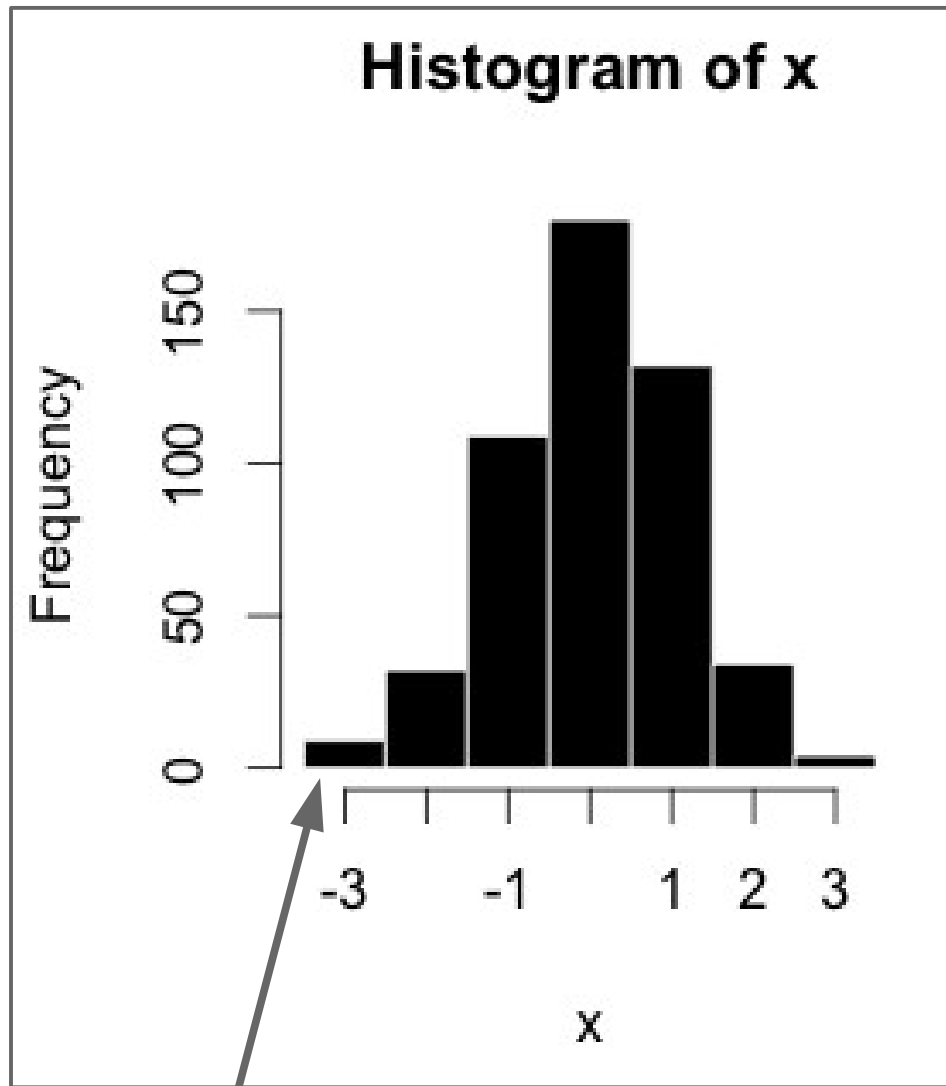
- on the x-axis, we have observations of some quantity
in our case, x = stellar ages

- the y-axis counts the number of occurrences of x

So, if out of 91 stars (x), we have 12 stars with an age of 10 Gyr, the y value for $x = 10$ will be 12.

The number of bars corresponds to the number of “bins,” in this case, 7. The choice of bin number (or bin size) can have a noticeable effect on the shape of a distribution—this is notorious weakness of histograms.

Step 5: Histogram and bins



- on the x-axis, we have observations of some quantity
in our case, x = stellar ages

- the y-axis counts the number of occurrences of x

So, if out of 91 stars (x), we have 12 stars with an age of 10 Gyr, the y value for $x = 10$ will be 12.

- ▼ Step 5: create a histogram of the ages

```
[ ] histogram = np.histogram(Joyce_ages)
```

- ▼ Now, grab the bins from the histogram we have created

```
[ ] bins = histogram[1]
    ## you may print the bins by uncommenting the line below
    #print("bins: ", bins)
```

The number of bars corresponds to the number of “bins,” in this case, 7. The choice of bin number (or bin size) can have a noticeable effect on the shape of a distribution—this is notorious weakness of histograms.

Step 6: Creating a Gaussian model for our data

- we now have bins, mu, and sigma defined for our data (mu = Jmu, sigma = Jsigma)
- we can think of the bins as serving the role of x in the Gaussian function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

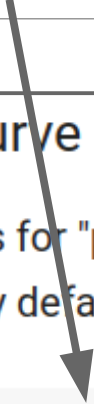
Step 6: Creating a Gaussian model for our data

- we now have bins, mu, and sigma defined for our data (mu = Jmu, sigma = Jsigma)
- we use the scipy.stats function **norm.pdf()** to create our model with this information

{x} ▾ Step 6: Create a curve defined by μ, σ

□ "pdf" in norm.pdf stands for "probability density function," and it is normalized such that its area is 1 by default

```
[ ] normalized_gaussian_pdf = norm.pdf(bins, Jmu, Jsigma)
```



Step 6: Creating a Gaussian model for our data

- we now have bins, mu, and sigma defined for our data (mu = Jmu, sigma = Jsigma)
- we use the scipy.stats function **norm.pdf()** to create our model with this information

{x}

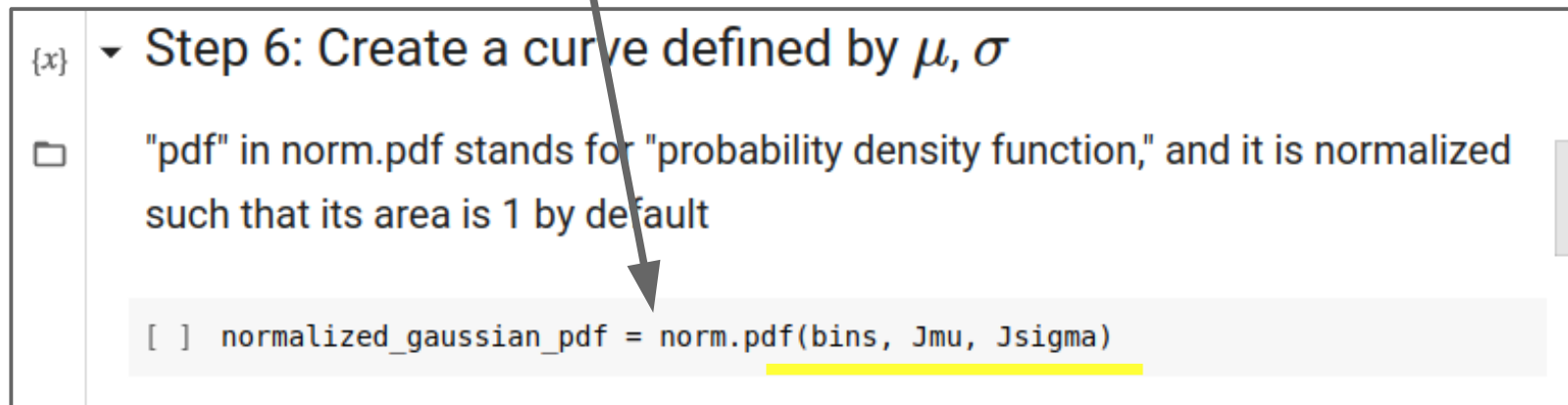
▼ Step 6: Create a curve defined by μ, σ

📁

"pdf" in norm.pdf stands for "probability density function," and it is normalized such that its area is 1 by default

[]

```
normalized_gaussian_pdf = norm.pdf(bins, Jmu, Jsigma)
```



However (!!)

this will generate a Gaussian whose integral is equal to 1, by definition: AKA “normalized”

Step 6: Creating a Gaussian model for our data

{x} ▾ Step 6: Create a curve defined by μ, σ

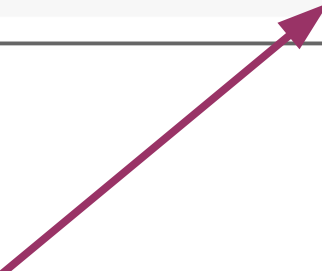
📁 "pdf" in norm.pdf stands for "probability density function," and it is normalized such that its area is 1 by default

```
[ ] normalized_gaussian_pdf = norm.pdf(bins, Jmu, Jsigma)
```

Now, rescale the curve so that it fits the size of our data. There are

- ▾ 91 age measurements, so `len(Joyce_ages) = 91`. We multiply our normalized Gaussian by this value

```
[ ] gaussian_pdf= normalized_gaussian_pdf*len(Joyce_ages)
```



So, we rescale the normalized Gaussian pdf so that its integral (~ sum over discrete bins) is equal to our data size (91 stellar ages)

Step 7: Graphically compare data and model

Histogram of
the Joyce age
measurements

Gaussian
model vs bins

Step 7: Plot our histogram and the Gaussian curve we have fit to it

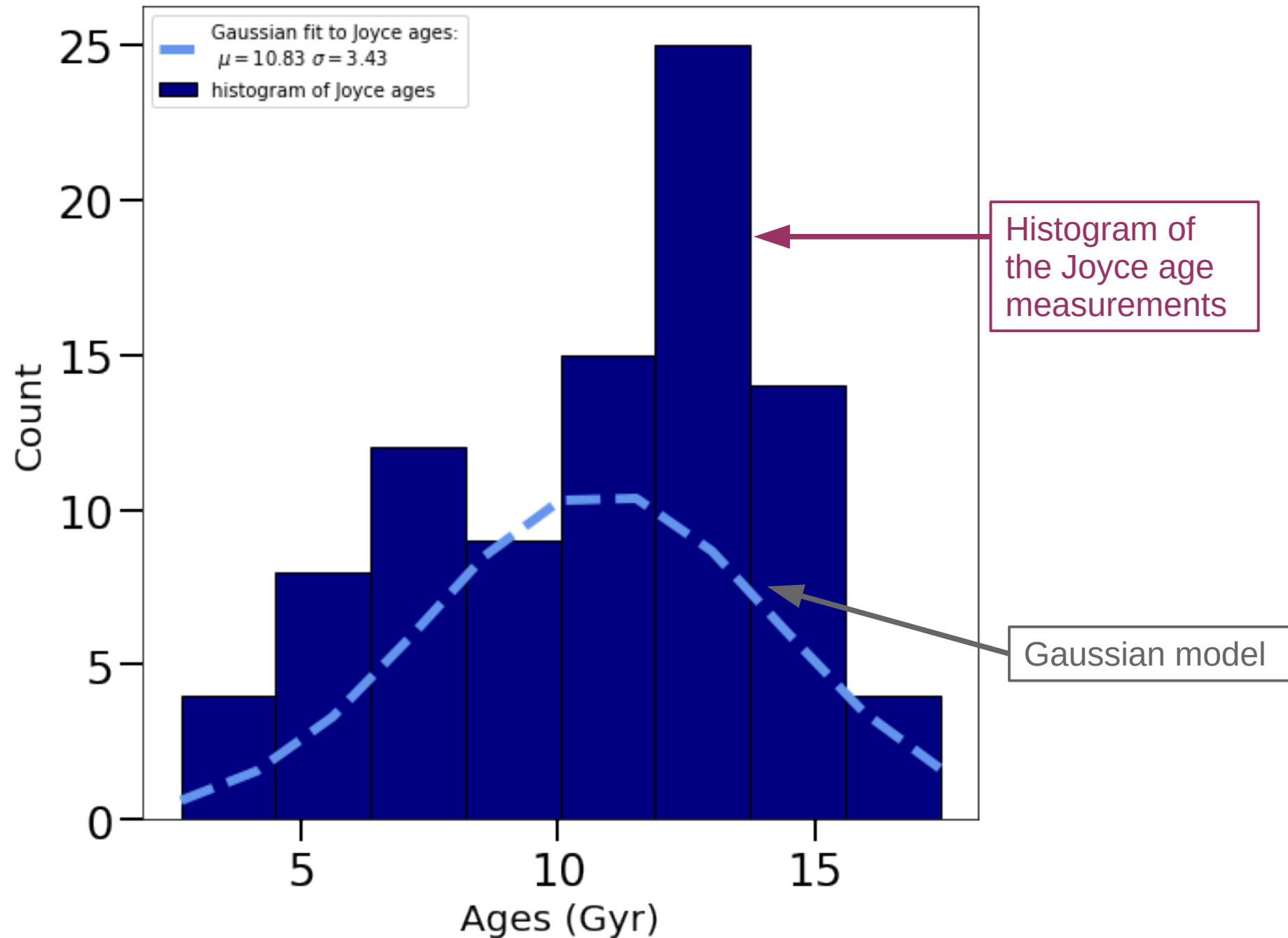
```
▶ ## initiate the figure
fig, ax = plt.subplots(figsize = (8,8))
set_fig(ax)

## this is the histogram
plt.hist(Joyce_ages, bins="auto", color='navy', edgecolor='black', label='histogram of Joyce ages')

## this is the Gaussian curve
plt.plot(bins, gaussian_pdf, \
        '--', color='cornflowerblue', linewidth=5, \
        label='Gaussian fit to Joyce ages:\n  $\mu$ = + %.2f%Jmu + '  $\sigma$ = + %.2f%Jsigma )

## these lines are plot bookkeeping
plt.xlabel('Ages (Gyr)', fontsize=20)
plt.ylabel('Count', fontsize=20)
plt.legend(loc=2)
plt.show()
plt.close()
```

Step 7: Graphically compare data and model



Pause – how good is this fit?

Pause – how good is this fit?

Not great!

Kernel Density Estimation



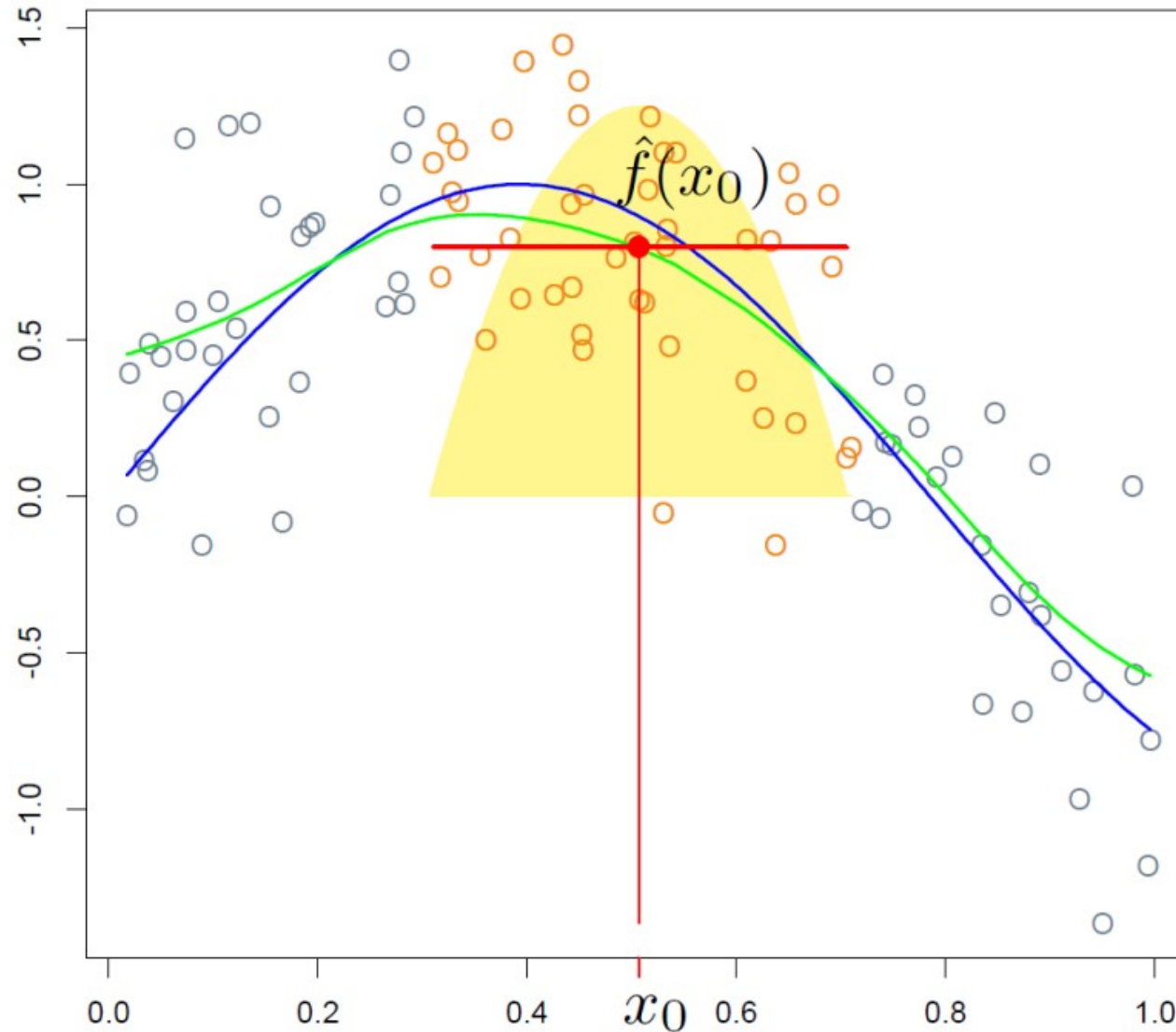
Kernel Density Estimation



- *kernel smoothing* is a statistical technique used to infer a function based on the local clustering (or density) of observed data



Kernel Density Estimation



We may think of “sliding” the yellow region—the kernel—across the data to generate local weighted averages of the data that combine to form the smoothed curve

Kernel Density Estimation



- *kernel smoothing* is a statistical technique used to infer a function based on the local clustering (or density) of observed data
- the scale over which the data are “grouped” and averaged determines the “smoothness” of the kernel. This is parameterized by the *bandwidth*, sometimes denoted with h



Kernel Density Estimation



- *kernel smoothing* is a statistical technique used to infer a function based on the local clustering (or density) of observed data
- the scale over which the data are “grouped” and averaged determines the “smoothness” of the kernel. This is parameterized by the *bandwidth*, sometimes denoted with h
- Kernel methods are *non-parametric* estimators, meaning they require no constraints from theory. Neural networks are another type of non-parametric technique



Kernel Density Estimation

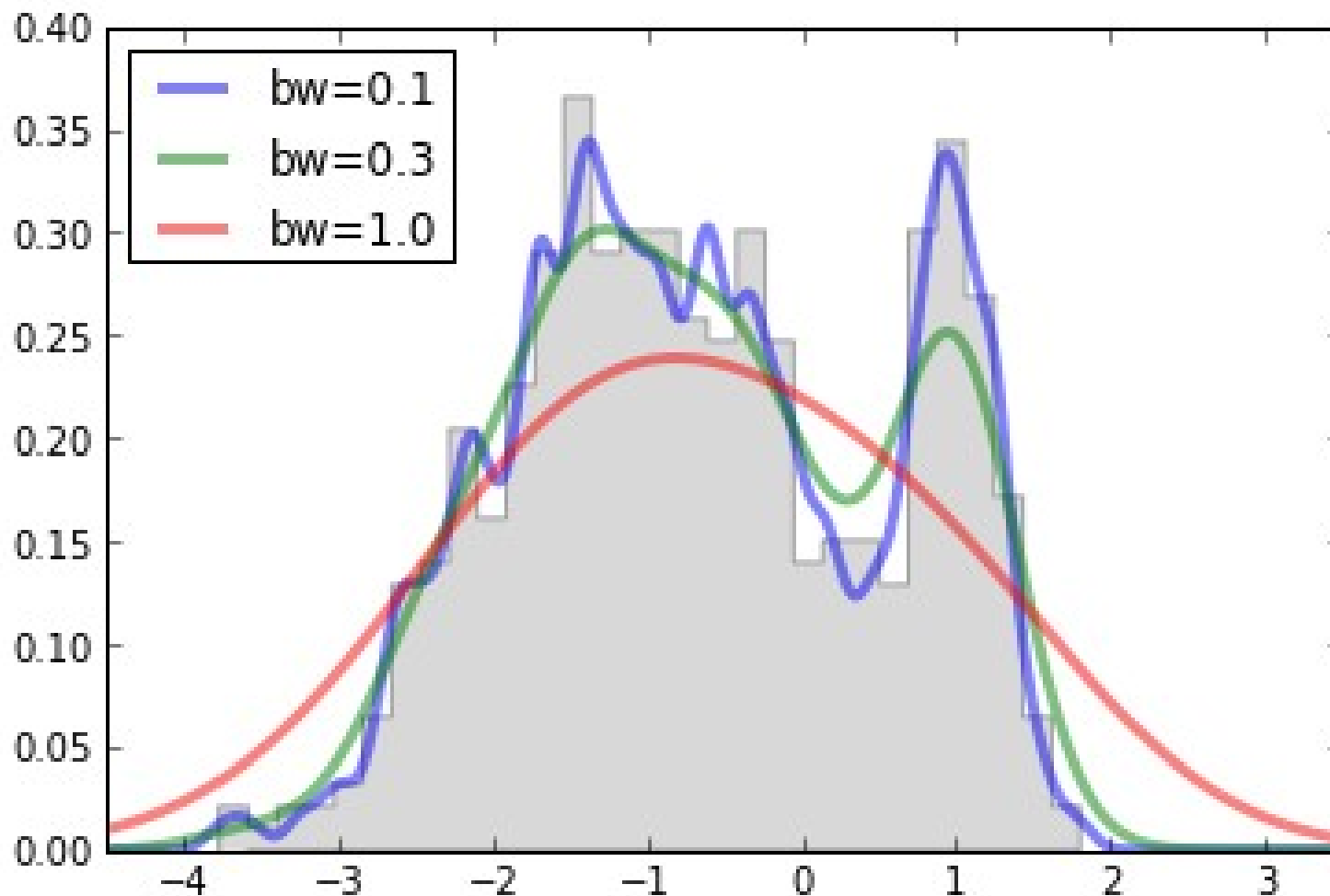


- *kernel smoothing* is a statistical technique used to infer a function based on the local clustering (or density) of observed data
- the scale over which the data are “grouped” and averaged determines the “smoothness” of the kernel. This is parameterized by the *bandwidth*, sometimes denoted with h
- Kernel methods are *non-parametric* estimators, meaning they require no constraints from theory. Neural networks are another type of non-parametric technique
- ***Kernel Density Estimation*** is the application of kernel smoothing to estimate the probability density function of a (continuous) random variable (e.g. stellar age) based on finite observed data (91 specific measurements of stellar age)



Kernel Density Estimation

KDEs of some observed data (grey histogram) with three different bandwidths (bw)



Note that using a bandwidth of 1.0 results in a function that fails to capture some features of the observed distribution (**underfit**), whereas use of $h = 0.1$ **overfits** the data

There is an optimal way to pick a bandwidth, but we won't get into that here

Kernel Density Estimation

There is plenty of sophisticated theory behind this, but from the Python perspective, it's just another model

A cluster of yellow corn cobs is positioned in the background, partially obscured by the text. The cobs are bright yellow and appear to be scattered or piled together.

Step 8: KDE for Joyce ages

▼ Step 8: Try a Kernel Density Estimate (KDE) instead

Create the kde model for the stellar ages

```
▶ kde_model = stats.gaussian_kde(Joyce_ages)
```

▼ to make the model smoother, we can increase the resolution of the x-axis

the line below subdivides the age range into 1000 equally spaced values. The age range is the minimum age measurement, `min(Joyce_ages)`, to the maximum age measurement, `max(Joyce_ages)`. These correspond to about 2 Gyr (billion years) and 17 Gyr, respectively

```
[ ] age_x_values = np.linspace(min(Joyce_ages), max(Joyce_ages), 1000)
```

▼ the following line evaluates the kde_model function we made at the beginning of Step 8 over the smoother array of x values defined above.

```
[ ] kde = kde_model(age_x_values)
```

▼ once again, the model is normalized to 1, so we must rescale it by the number of age measurements

```
[ ] ## scale the kde by the number of stellar ages in our sample (91)  
    scaled_kde = kde*len(Joyce_ages)
```

Step 9: Compare KDE graphically

Step 9: Now, add our KDE model curve to the histogram plot from Step 7

```
fig, ax = plt.subplots(figsize = (8,8))
set_fig(ax)

## histogram from earlier
plt.hist(Joyce_ages, bins="auto", color= 'navy', edgecolor='black', label='histogram of Joyce ages')

## Gaussian fit from earlier
plt.plot(bins, gaussian_pdf,\
        '--', color='cornflowerblue', linewidth=5,\
        label='Gaussian fit to Joyce ages:\n  $\mu$ =$'+ "%.2f"%Jmu + '  $\sigma$ =$'+ "%.2f"%Jsigma)

## NEW: add the KDE to the plot
plt.plot(age_x_values, scaled_kde,\
        linewidth=5, linestyle='--', color='lightblue',\
        label='KDE of Joyce age distribution')

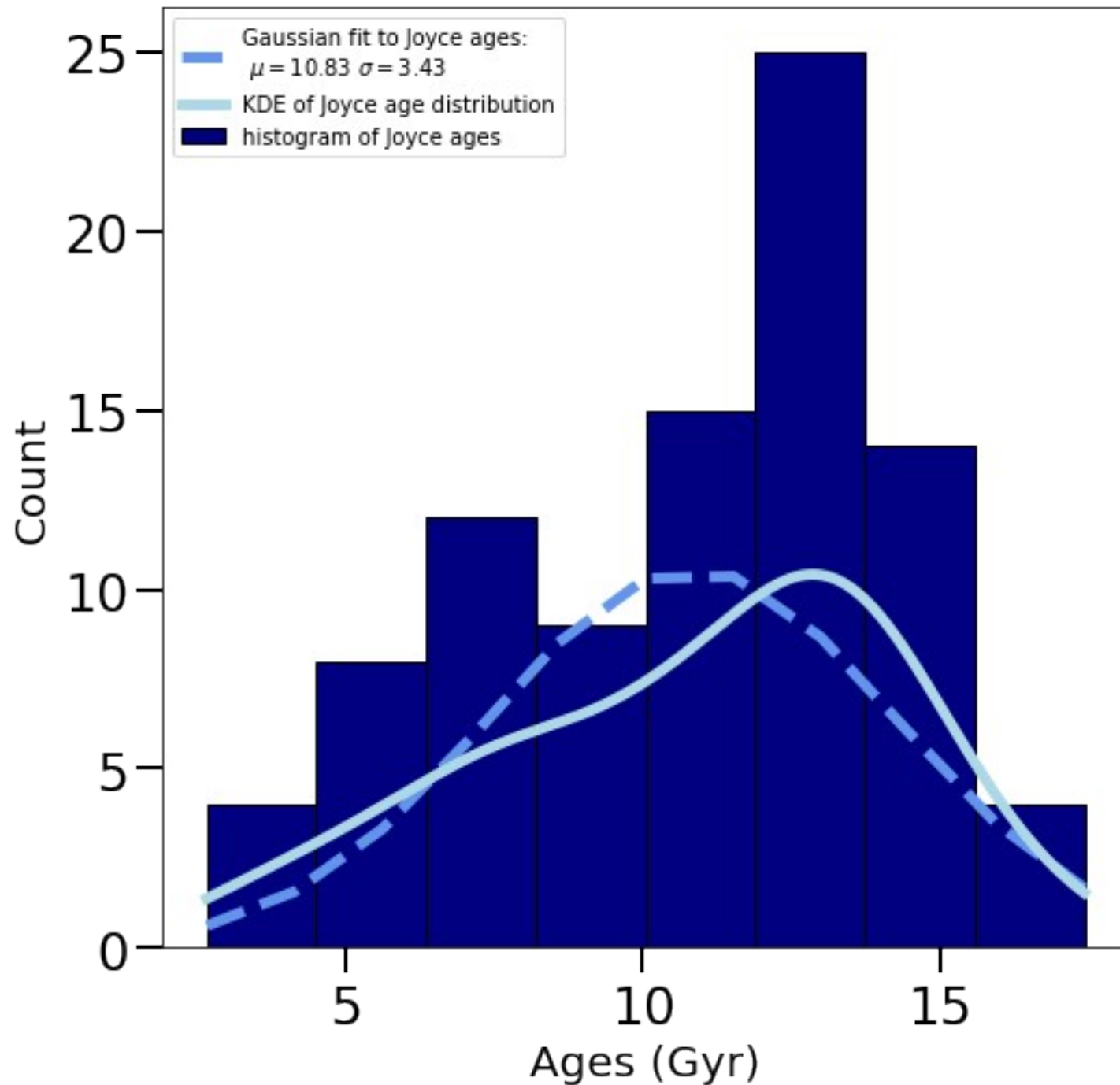
plt.xlabel('Ages (Gyr)', fontsize=20)
plt.ylabel('Count', fontsize=20)
plt.legend(loc=2)
plt.show()
plt.close()
```

Joyce age
histogram

Gaussian model

NEW: KDE model

Step 9: Compare KDE graphically



Try Exercises 1 and 2

Consult the solutions if you are stuck

If you finish that, read **PART 2: KDE Resampling and Generative Models** and try the exercises