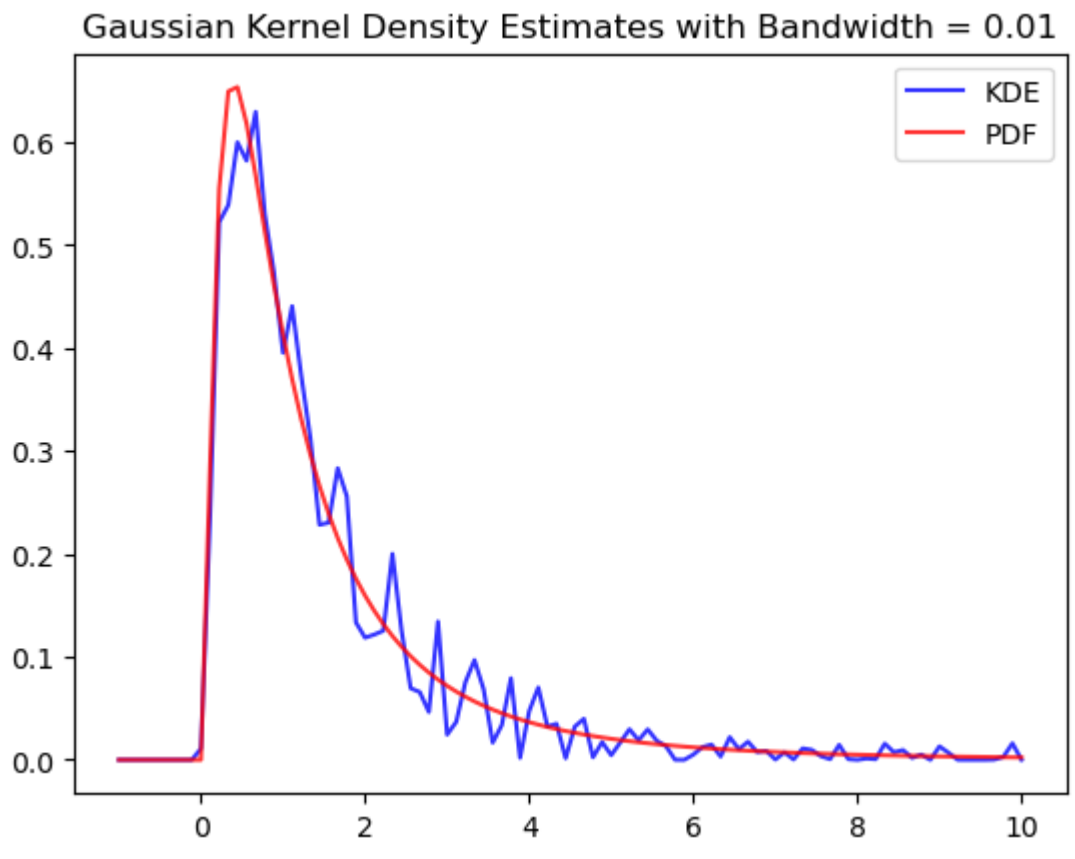
Homework 2

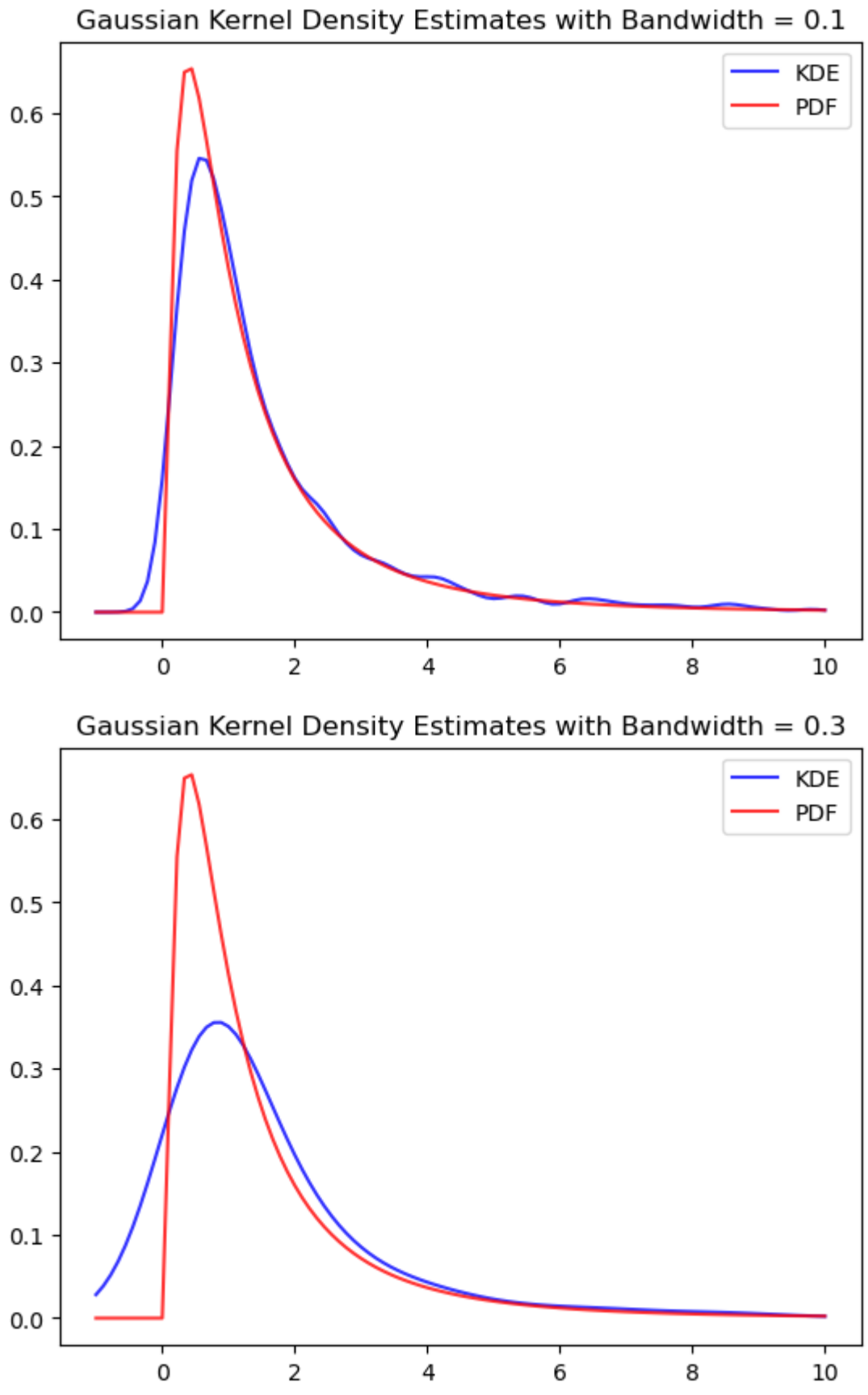
It can be problematic to use standard kernel density estimation to approximate a density with bounded support. Demonstrate this as follows:

- Simulate lognormal data.
- Fit Gaussian kernel density estimates using `scipy.stats.gaussian_kde`.
- Display three plots showing the tradeoff between fit at the support boundary and overall fit.

Explain why the phenomenon demonstrated in the plots occurs.

In [95]:
```
# CODE
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import lognorm
import scipy.stats as stats

#Simluate lognormal data
x = np.random.lognormal(0, 1, 1000)
#Fit Gaussian kernel density estimates with different bandwidth
k_1 = stats.gaussian_kde(x, bw_method = 0.01)
k_2 = stats.gaussian_kde(x, bw_method = 0.1)
k_3 = stats.gaussian_kde(x, bw_method = 0.3)
a = np.linspace(-1, 10, 100)
mu, sigma, n = lognorm.fit(x)

#plot the 3 results
y_1 = k_1.evaluate(a)
plt.plot(a, y_1, color='b',alpha=0.8, label='KDE')
plt.plot(a, stats.lognorm.pdf(a,mu,sigma), color='r',alpha=0.8, label='PDF')
plt.title("Gaussian Kernel Density Estimates with Bandwidth = 0.01")
plt.legend()
plt.show()

y_2 = k_2.evaluate(a)
plt.plot(a, y_2, color='b', alpha=0.8, label='KDE')
plt.plot(a, stats.lognorm.pdf(a,mu,sigma), color='r',alpha=0.8, label='PDF')
plt.title("Gaussian Kernel Density Estimates with Bandwidth = 0.1")
plt.legend()
plt.show()

y_3 = k_3.evaluate(a)
plt.plot(a, y_3, color='b',alpha=0.8, label='KDE')
plt.plot(a, stats.lognorm.pdf(a,mu,sigma), color='r',alpha=0.8, label='PDF')
plt.title("Gaussian Kernel Density Estimates with Bandwidth = 0.3")
plt.legend()
plt.show()
```





Explanation:

Fromt he three plots, we can see that different bandwidth will deliever different results on the fitting of the support boundary. In order to have a better fitting, we have to choose the appropriate bandwidth which here is around 0.1. As we can see that a bandwidth of 0.01 is way too small that leads to an overfitting of the pdf and a bandwidth of 0.3 is too large that causes an underfitting of the pdf.