Problem Set #1

MACS 30250, Dr. Evans
Due Monday, May. 4 at 1:30pm

1. **1D kernel density estimator (5 points).** The COVIDincubation.txt is a data file that includes ? observations on the following two variables: `age`, `symp_days` (days symptomatic). This is a synthetic dataset that is meant to mimic the underlying data from [citation]. Each observation represents an individual who was known to have tested positive for the COVID-19 virus. The `symp_days` variable represents the incubation period for each individual, or the number of days until symptoms were manifest.

   (a) Create three histograms, each of `symp_days` (Incubation period, days to symptomatic). The first one is the overall histogram. Let each histgram have 15 bins over the range of days from zero to the maximum in the data. Let the first histogram be for all the data. Let the second histogram be for individuals of `age`< 40, and let the third histogram be for individuals `age`> 40.

   (b) Fit a Gaussian KDE to each histogram. Use LOOCV as in the Vander-Plas notebook to choose an optimal bandwith. Plot each of the KDE distributions in one plot with a legend that shows which is which.

2. **2D kernel density estimator (5 points).** The data `BQmat_orig.txt` is a 78×7 matrix of percentages representing the values of a two-dimensional histogram of the percent of the U.S. population that receives all the bequests (inheritances) by a recipient's age (ages 18 to 95, rows) and by a recipient's lifetime income group (7 categories, columns). The seven lifetime income groups are percentiles. Let $prcntl_j$ be the percent of the population in lifetime income group j. The lifetime income groups in the $J = 7$ columns of the `BQmat_orig.txt` data are the following.

$$\boldsymbol{prcntl} = \big[0.25, 0.25, 0.20, 0.10, 0.10, 0.09, 0.01\big], \quad \text{such that} \quad \sum_{j=1}^{7} prcntl_j = 1$$

   You can read this file into memory using the `numpy.loadtxt` function.

```
bq_data = np.loadtxt('BQmat_orig.txt', delimiter=',')
```

   So the $[11, 5]$-th element of the `bq_data` matrix represents the percent of total bequests (inheritances) received by age-28 and lifetime income group $j = 5$ (80th to 90th percentile of lifetime income).

(a) Read in the bequests data as a 78×7 `NumPy` array. Plot the 2D empirical histogram of these data as a 3D surface plot with age and income group on the x-axis and y-axis and the histogram density on the z-axis using a 3D surface plot tool (not a 3D bar histogram tool). Make sure that the axes are labeled correctly. And make sure that your 3D histogram is presented from a perspective that allows a viewer to see that data (don't let the data be hidden by a poor angle of the plot.)

(b) Fit a bivariate kernel density estimator to the data. Use a Gaussian kernel. Choose a bandwidth parameter λ that you think is best. Justify your choice of that parameter. Plot the surface of your chosen kernel density estimator. Make sure that the axes are labeled correctly. And make sure that your 3D histogram is presented from a perspective that allows a viewer to see that data. What is the estimated density for bequest recipients who are age 61 in the 6th lifetime income category ($j = 6$, 90th to 99th percentile).