**DSSC 221 – Probability and Statistics**
**Lab 8**

Data file "data.txt" is in LMS. The data set contains the total number of residential acres in each planning zone within the City of Dublin, California, based on 2005 ABAG projections. The data set consists of approximately 2650 observations, and we consider this to be our complete data set.

1. Data preparation:

   Import the data file as pandas dataframe (df).

2. Sampling distribution & central limit theorem
   a. Plot the histogram of df (hint: try 50 bins). How do you think this data is distributed?
   b. Find the mean and variance of the data.
   c. Define a function called *CLT*. The code will randomly draw samples of size "nSample", up to "nSampleSets" times. (*For instance if nSample=10 and nSampleSets=100, we would randomly draw a set of 10 observations 100 times).*

      - Complete the code by implementing the followings:

      i. **Challenge Question (Extra pts.) -** "actVarSampleMean", which refers to the empirical sample variance of the means. Please enter in the correct code to calculate that.

      ii. **Challenge Question (Extra pts.) -** Just below you will find "theorVarSampleMean", which refers to the theoretical sample variance. Please enter the correct equation to calculate that.

      - Then, the code automatically finds the means of the samples and plots them.

      iii. **Challenge Question (Extra pts.) -** Run the code with nSample=10 and nSampleSets=1000, and compare the result with nSample=10 and nSampleSets=100. What do you observe?

      iv. **Challenge Question (Extra pts.) -** Run the code with nSample=100 and nSampleSets=1000, and compare the result with nSample=10 and nSampleSets=1000. What do you observe?

      After each run, comment on the shape of each resulting histogram of the means. Also compare "actVarSampleMean" versus "theorVarSampleMean".