1. Select a dataset from one of these sources:

http://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets
http://archive.ics.uci.edu/ml/datasets.html
https://www.kaggle.com/datasets

2. Include explanations of your choice in the presentation (why are you interested in this particular data set or perhaps why do you think this is important; what were you hoping to discover and did your discovery agree with your expectations)

3. (Visualization) For each of the selected variables:
   a) represent these data sets in frequency tables;
   b) display histogram; relative frequency line graph; cumulative relative frequency plot.
   c) Are any of the histograms in (b) approximately normal?

4. (Modeling) For each of the selected variables assume that it follows one of the theoretical distributions (choose 2 from Continuous or Discrete as appropriate for your variable)
   a) Set up each distribution, using the appropriate information from your data.
   b) Must your data fit one of the above distributions? Explain why or why not.
   c) Could the data fit 2 or 3 of the above distributions (at the same time)? Explain.
   d) Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.
   e) Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

6. Compute
   a) sample mean, median, mode, variance, standard deviation;
   b) Determine the proportion of the data values that lies within 1.5 IQR
   c) Assuming population variance is the same as sample variance, construct 95% confidence interval for the population mean based on the entire data set on the portion determined in b). Are they different? Explain.
   d) Assuming population variance is unknown, construct 95% confidence interval for the population mean. Is it different from CI computed in part c)? Explain.

7. (CLT) Use a random number generator to pick $N$ samples of size $n$ from your original data. Use $N = 10, 100, 500$ and $n = 2,5,10$. For each choice of $N$ and $n$:
   a) Compute the average $\bar{X}$.
   b) Base this on the mean and standard deviation from your original data, state the approximate theoretical distribution of $\bar{X}$.
   c) Construct a histogram displaying your data.
   d) Draw the graph of the theoretical distribution of $\bar{X}$ and compare the relative frequencies to the probabilities. Are the values close?
   e) Does it appear that the data of averages fit the distribution of $\bar{X}$ well?

What happened to the shape and distribution when you averaged your data? In theory, what should have happened? In theory, would it always happen? Why or why not?

8. Choose a pair of variables, which demonstrate a relationship via correlation.
   a) Draw a scatter diagram relating these variables
   b) Determine the sample correlation coefficient
   c) What conclusion(s) can you draw from (a) and (b).

9. Stratify one of the variables into categories (or use a categorical variable from your data set) and construct boxplot for each category. Discuss what (if anything) you have observed from this type of visualization technique.

**Summarize in slides and upload to blackboard (pdf only).**