DASC 500 Homework 1

1. Using the CommuteTimes file provided, calculate the:
   a. *Mean* for the commute times
   b. *Median* for the commute times
   c. *Mode* for the commute times
   d. *Population variance* and *population standard deviation* for the commute times
   e. *Third quartile* for the commute times

2. Using the JoblessRate file provided
   a. Calculate the *mean* for:
      i. jobless rate %
      ii. delinquent loan %
   b. Calculate the *median* for:
      i. jobless rate %
      ii. delinquent loan %
   c. Calculate the *sample variance* and *sample standard deviation* for:
      i. jobless rate %
      ii. delinquent loan %
   d. Calculate the (Pearson) *correlation coefficient* between the jobless rate % and delinquent loan% data.
   e. Develop an individual *histogram* for each data set listed below that best portrays its distribution:
      i. jobless rate %
      ii. delinquent loan %.

3. In your own words, describe each phase of the CRISP-DM Process Methodology.

4. List the generic tasks included in the CRISP-DM Business Understanding Phase.

5. List the generic tasks included in the CRISP-DM Data Understanding Phase.

6. Search the web or a preferred source and provide an example of poor or misleading data visualization. Explain why you think it is a poor or misleading example.

Extra Credit Question on next page

Extra Credit Problem (worth 20%)

Use the Python interactive binning function and the recommendations given in either Histograms: A Useful Data Analysis Visualization or the Ultimate Guide to Binning to:

- Identify an initial bin width suggested for the Commute Times data set
- Select the bin width that best represents the distribution of the Commute Times data
- Present the histogram that best represents the distribution of the Commute Times data

You may use Excel to answer this question if you're still warming up to Python.