**Homework 1 [PDS Fall 2013]**

**Practical Data Science, Fall 2013**

**Homework 1: Preliminary exercises in Data Science**

**Modeling Example**

Try to code this example (everything after source) using nearest neighbors to try and classify examples correctly on the standard iris dataset in an Ipython notebook. An explanation of the data is here. Try to modify the number of nearest neighbors considered and note how the regions change.



Distance remains fairly similar, with slight deviations at all stages.

Uniform remains fairly similar until it hits 100, 200 and 100. At 100, the green section becomes larger, at 200 and 1000, red dominates the entire graph.

**Income Prediction 1.**

1. Download marketing.data located here. This data set is described here.

2. How many lines are in the file? Hint: use the wc unix utility.

wc -l marketingdata.txt

8994 marketingdata.txt

3. Notice that many lines have some fields unavailable (NA). Remove any lines without complete data. Hint: use the grep unix utility. How many lines remain?

grep -v NA marketingdata.txt > marketingdatap3.txt

wc -l marketingdatap3.txt

6877 marketingdatap3.txt

4. The fifth column corresponds to education level. What is the most common education level?



4.0

According to the information, that corresponds to "1 to 3 years of college".

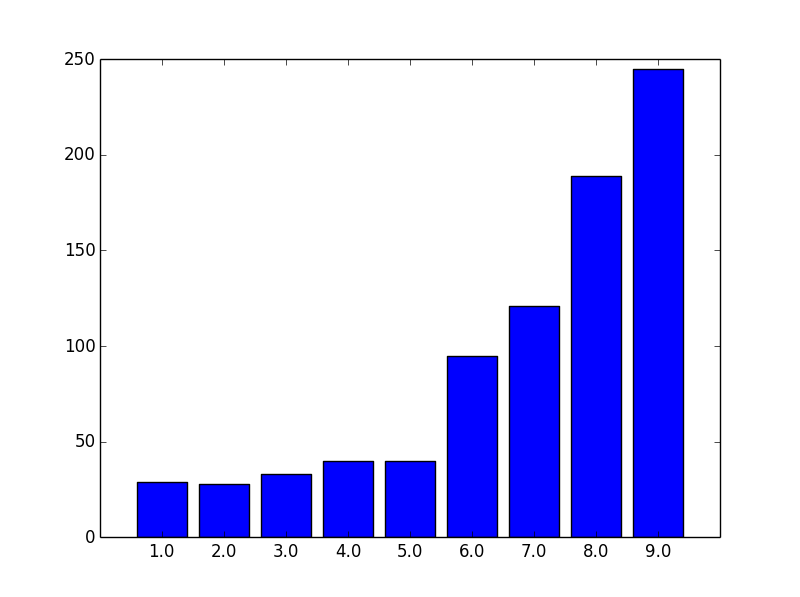
5. What is the income distribution for households with some graduate school? (hint: use a python dict data structure to store income level counts)



This shows that income distribution for households with some graduate school typically earn along the higher end.

This can then be visualized.





Which very clearly shows that households with some graduate school typically earn along the higher end of the income spectrum.