CS 7280 - Spring 2015 Olga Vitek

Homework 1

Each part of the problems 5 points
Due on Blackboard before 11:45am on Thursday January 21.

1. Download a dataset of your choice from this website `https://r-dir.com/reference/datasets.html`. Every dataset is acceptable, as long as it is structured as an array (as opposed to, e.g., a network). Describe the nature of the data in text, and then provide basic summaries in R. Use Sweave or Markdown to prepare the report (see examples on the course website). *Note:* If a dataset is too large, it is ok to take a subset.

   (a) How many observations are in the dataset? How many features?

   (b) Which features are continuous? Categorical? Factors?

   (c) Provide summaries of each feature (i.e., quartiles for continuous features, counts for categorical/factors)

   (d) Make univariate plots for every feature, and bivariate plots for pairs of features that may be important.

2. Use the following code to select a subset of 50 diamonds from the `diamonds` dataset in the library *ggplot2*, and plot 2 histograms.

```
set.seed(123)
index <- sample(1:nrow(diamonds), 50) # try a subset first
diamonds2 <- diamonds[index,]
hist(diamonds2$price)
hist(diamonds2$price, breaks=quantile(diamonds2$price))
```

   (a) Manually repeat the calculations behind the two histograms, and verify that your manual calculation corresponds to R. (See example on the course website).

   (b) Use the two histograms to report the proportion of diamonds with price less than $4,000. Verify that the two histograms give the same result.