EDA Homework 1

Due: Friday, January 29 at 5pm

Submit exactly two files: (i) a PDF/HTML file with your write-up and graphs and (ii) a .r/.txt/.Rmd file with code to reproduce your graphs. Haoran will randomly check some of you to ensure your graphs can be reproduced from your code; if not, you'll be penalized.

The dataset `tips` (available in the `reshape` package) contains data on tipping. The variables are

- `total_bill` in dollars
- `tip` in dollars
- `sex`, male or female, of the bill payer
- `smoker`, whether or not there was a smoker in the party.
- `day` of the week.
- `time`, the time of day, either lunch or dinner.
- `size`, the number of people in the party.

Since the tip depends strongly no the total bill, the quantitative variable we'll study will be the percentage tipped. The lowest percentage tipped was 3.6% and the highest was 71%.

1. Use ggplot to draw *one* graph of the percentage tipped for the sample that enables you to see the center, spread, and shape of its distribution. Make sure your graph looks nice (meaningful titles and labels, sensible choices of any adjustable parameters). Describe, numerically or in words, the center, spread, and shape of the distribution. Is it normal?
2. Use a set of faceted plots to display the distribution of percentage tipped for each party size in a way that allows easy comparison. Are there clear differences between the distributions for different party sizes, or are they about the same? Or is it impossible to say? Explain.
3. Pick a measure of center for the percentage tipped distributions, and explain your choice. Use the `aggregate` function to find the center of percentage tipped for each party size, and display these using a dot plot. Without doing a formal statistical test, answer which (if any) differences in the centers for different party sizes look real and which can be reasonably explained by chance variation.