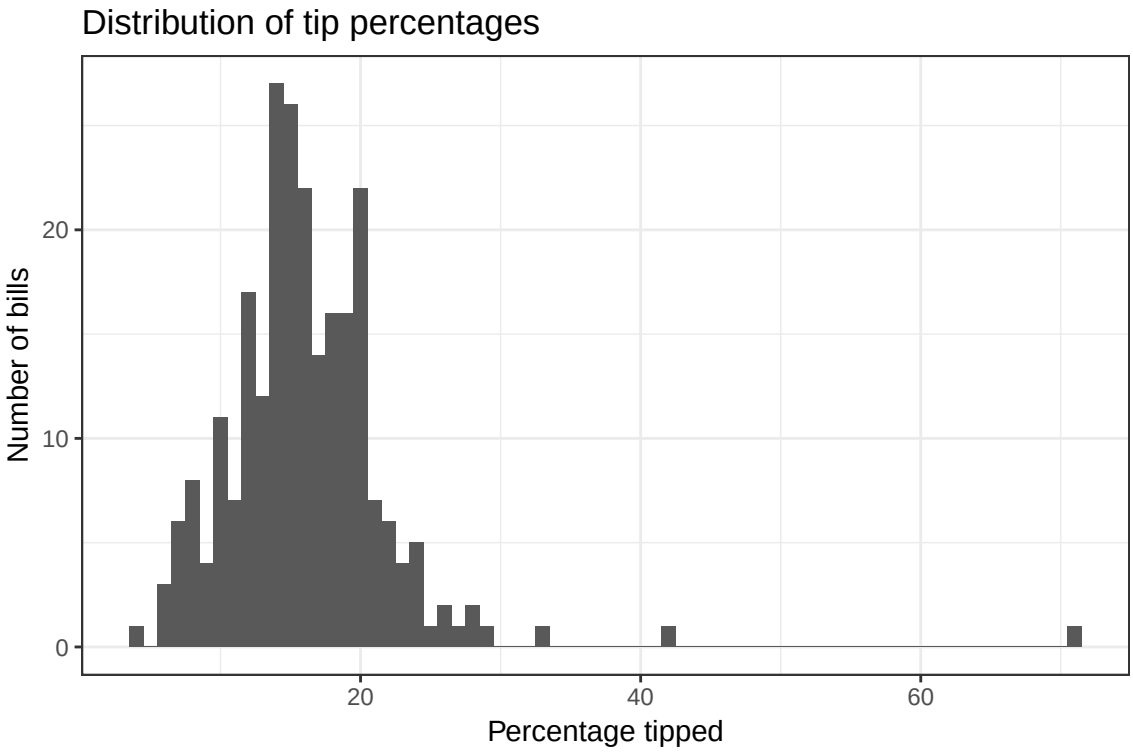
EDA Homework 1 Solutions

Question 1

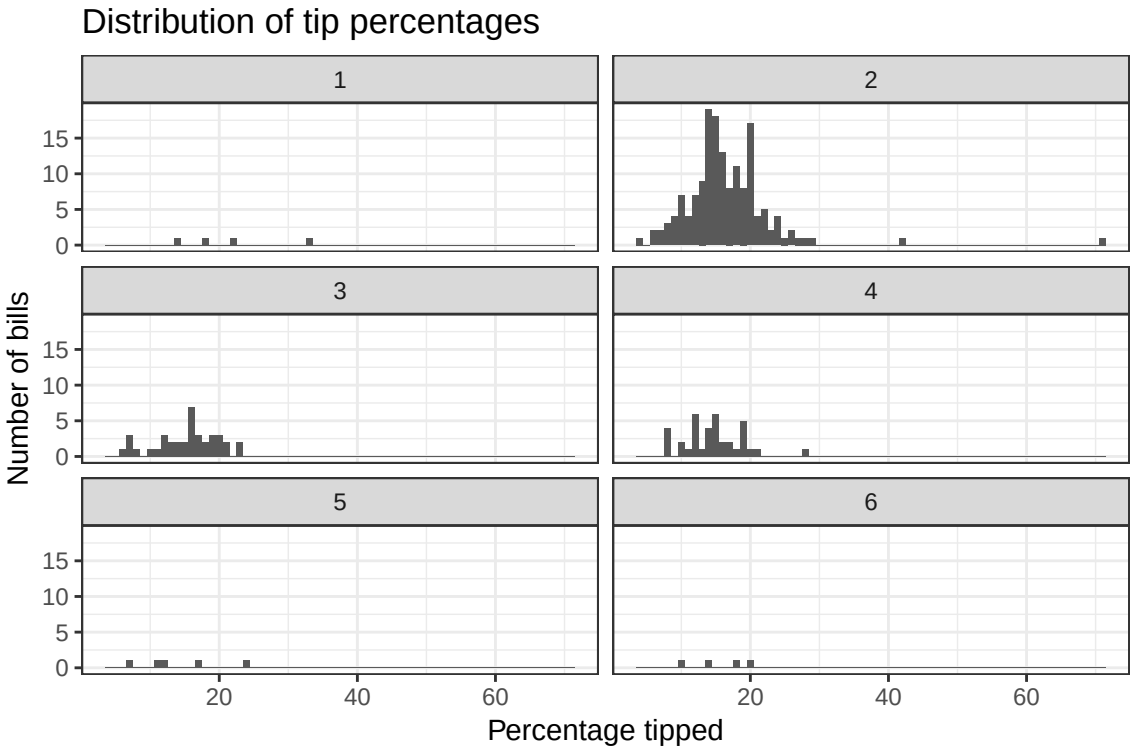
```
# Add tip percent to the data frame
tips$tip.percent = tips$tip / tips$total_bill * 100
library(ggplot2)
ggplot(tips, aes(x=tip.percent)) + geom_histogram(breaks=3.5:71.5) + xlab("Percentage tipped") +
```



Typically people tipped between 10% to 20%, with the usual rule-of-thumb 15% tip looking like it's at the center of the distribution. The distribution is slightly skewed, with a few extreme outliers (including a 71% tip.) Even ignoring the outliers, the distribution doesn't look normal – a close look suggests the distribution is "spiky" with multiple peaks, e.g. one around 15% and another around 20%.

Question 2

```
ggplot(tips, aes(x = tip.percent)) +
    geom_histogram(breaks=3.5:71.5) +
    facet_wrap(~size, ncol=2) +
    xlab("Percentage tipped") +
    ylab("Number of bills") +
    ggtitle("Distribution of tip percentages")
```

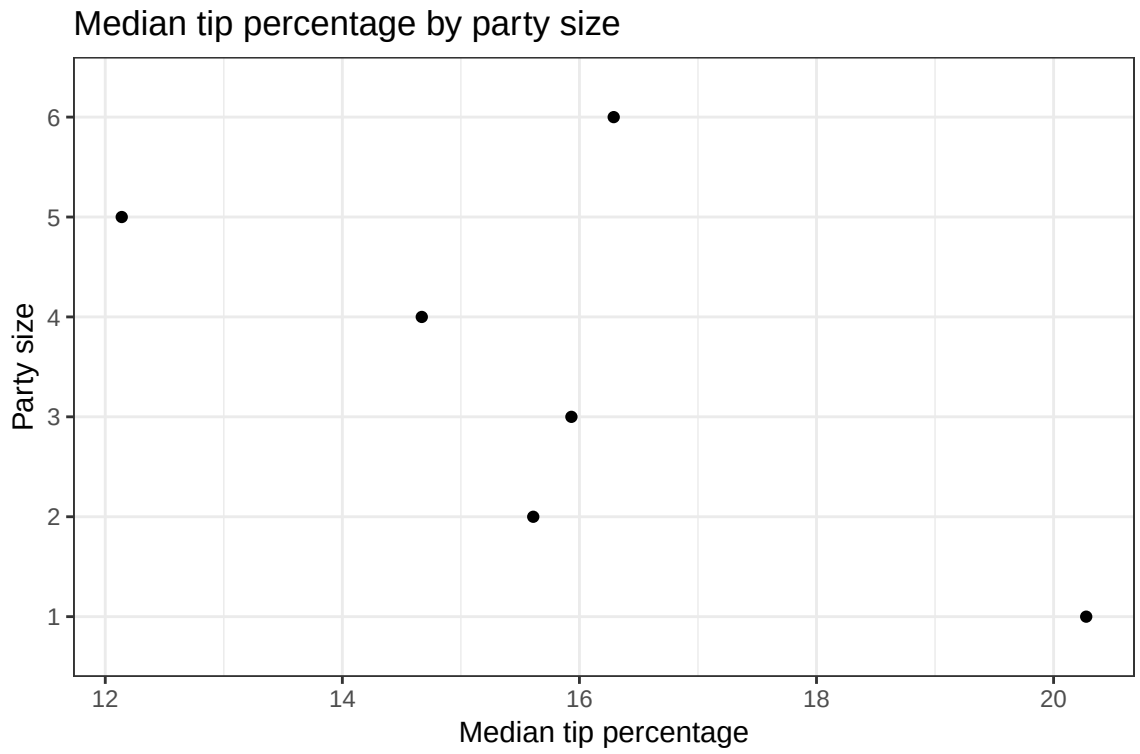


We have to ignore the results for party sizes of 1, 5, and 6 because of lack of data. Even for the the parties of size 3 and 4, the sample sizes are a bit small to tell if there really are differences in the shapes and spreads of the distributions. So really it's impossible to say if there really are meaningful differences of not. (This is often the case in EDA, as with any other kind of data analysis.)

Question 3

I'll use the median because it's generally a more meaningful measure of center for asymmetric distributions. However, you could also justify using the mean because it's more accurate (as a representation of a population) for small sample, although the extreme outliers might give you pause.

```
tip.medians = aggregate(tip.percent ~ size, FUN = median, data = tips)
ggplot(tip.medians, aes(y = factor(size), x = tip.percent)) +
    geom_point() +
    xlab("Median tip percentage") +
    ylab("Party size") +
    ggtitle("Median tip percentage by party size")
```



Again, we more or less ignore the party sizes of 1, 5, and 6 because of lack of data. The medians for party sizes from 2 to 4 are close enough (all between 14-16%) that all differences are easily attributable to chance.