EDA - Assignment1

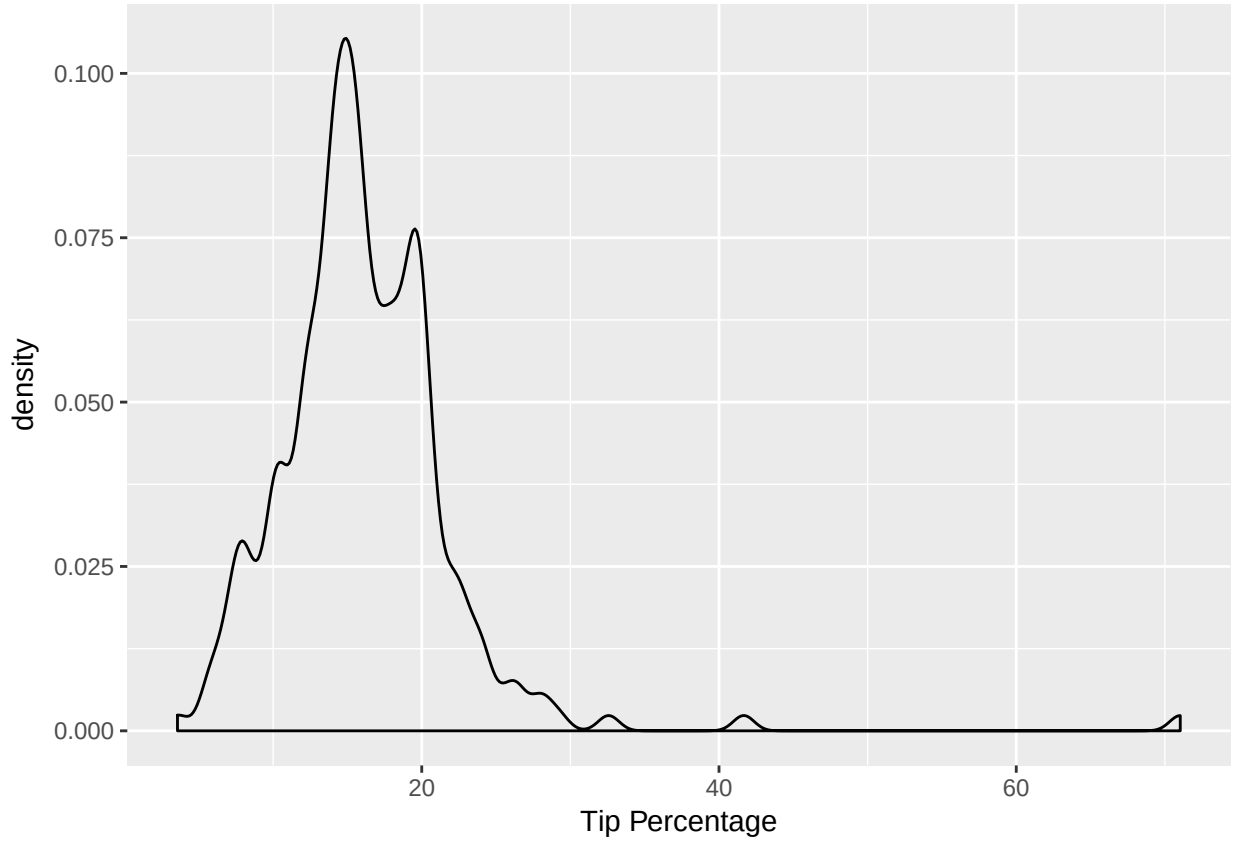
Jivitesh Poojary

January 16, 2017

Answer 1

Q1. Does the percentage tipped follow a normal distribution? If not, how does the data differ from a normal distribution? Include ONE graph drawn with ggplot to support your answer.

The below plot is a density plot of the PDF. An adjust value of less than 1 is used to make the plot less smooth, so that we can observe the distribution to a greater accuracy. From the distribution we can see that there are multiple peaks in the data, which violates the normality assumption. Similarly we can also observe that the distribution is skewed towards the right.



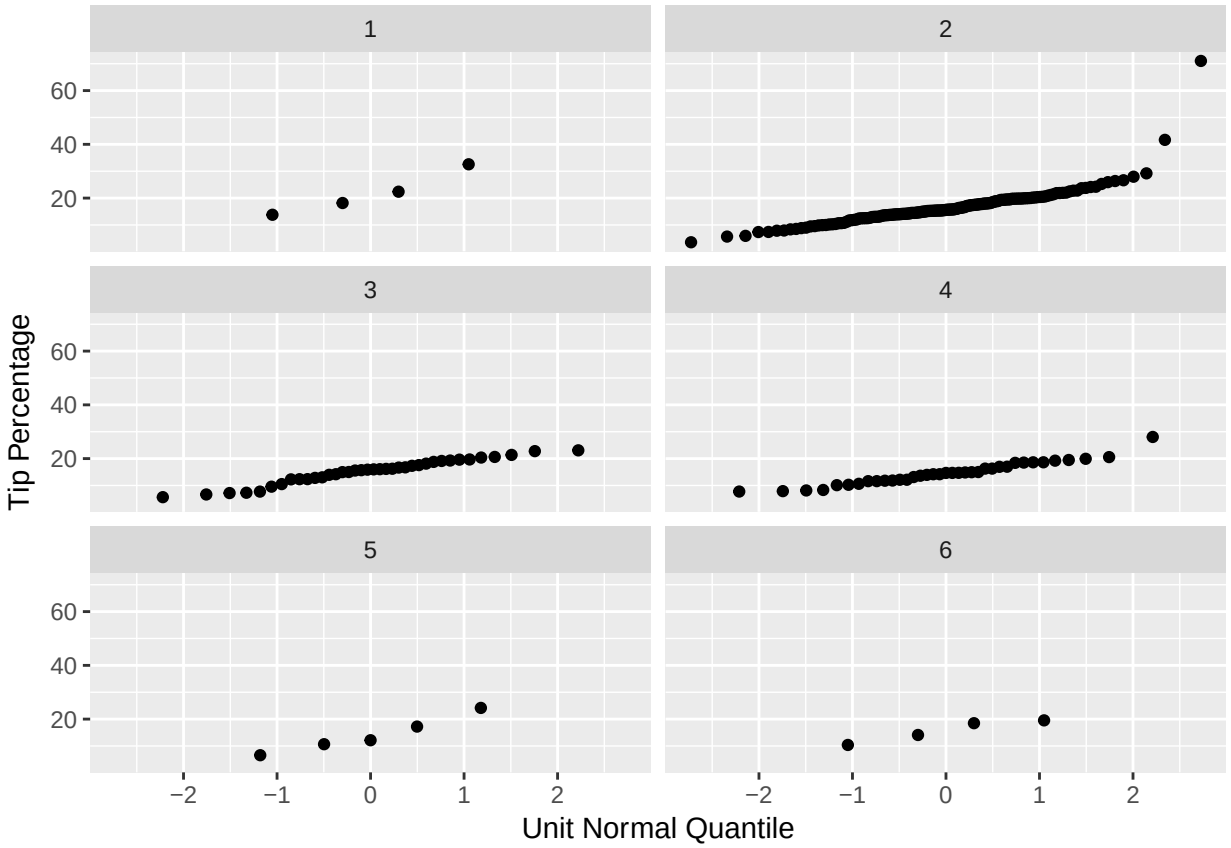
Answer 2

Q2. How does the distribution of tips change with party size? Include ONE plot to support your answer. (A set of faceted plots counts as one plot.).

There is insufficient data for the qq plots for party sizes 1,5 and 6. As a result we will not be able to confidently derive conclusions regarding the distribution. The highest tip percentage in all the party sizes is for party size 2, however on an average the tip percentage for party size 1 is high. Party size 2 also has the lowest tip percentage which makes think there are many outliers in the this distribution.

For the plots for party size 2,3 and 4 there is sufficient data, the plots for 3 and 4 show a good degree of normal distribution attributing to the straight line distribution of the points. For party size 2 the mean tip appears to be around 16, while for party size 3 it appears to be around 15 and for party size 4 it appears to be around 14.

The tip percentage are higher for lower party size. We can observe that mostly tip decreases with party size. There is a slight increase in the tip percentage when the party size reaches 6.



Answer 3

Q3. Using lm(), we can find that a linear model to predict percentage tipped from party size is Percent tipped = 18.4 - 0.92 * size Does it look like the model fits the data well? Include ONE plot of the residuals of this model to support your answer. (Again, a set of faceted plots counts as one plot.)

The fitted value plot is very flat compared to the residuals plot. The spread of the fitted value versus the spread of the residuals show that the spread is not comparable. The model fit does not account for a major proportion of the variation, most of the variation is captured in the residual plot.

Similarly, the R square value for the model is 0.02040887, this again shows that the current model does not capture the distribution of data.

The fitted - residual plot is used as we can visually compare the fit and residual distribution. A box plot or a fit with linear model plot would not have an impact in conveying the required information.

```
## [1] 0.02040887
```