Assignment1

Surbhi Paithankar, Apurva Gupta

January 15, 2018

Loading Data Into R

```
data = read.table("tips.txt",header = TRUE)
head(data)
```

```
##   total_bill  tip    sex smoker day   time size
## 1      16.99 1.01 Female     No Sun Dinner    2
## 2      10.34 1.66   Male     No Sun Dinner    3
## 3      21.01 3.50   Male     No Sun Dinner    3
## 4      23.68 3.31   Male     No Sun Dinner    2
## 5      24.59 3.61 Female     No Sun Dinner    4
## 6      25.29 4.71   Male     No Sun Dinner    4
```

Question 1

Calculating tip percentage

```
tips_percent = (data$tip/data$total_bill)*100
data<-cbind(data,tips_percent)
head(data)
```

```
##   total_bill  tip    sex smoker day   time size tips_percent
## 1      16.99 1.01 Female     No Sun Dinner    2     5.944673
## 2      10.34 1.66   Male     No Sun Dinner    3    16.054159
## 3      21.01 3.50   Male     No Sun Dinner    3    16.658734
## 4      23.68 3.31   Male     No Sun Dinner    2    13.978041
## 5      24.59 3.61 Female     No Sun Dinner    4    14.680765
## 6      25.29 4.71   Male     No Sun Dinner    4    18.623962
```

Importing ggplot Library and creating an object

```
library(ggplot2)
```

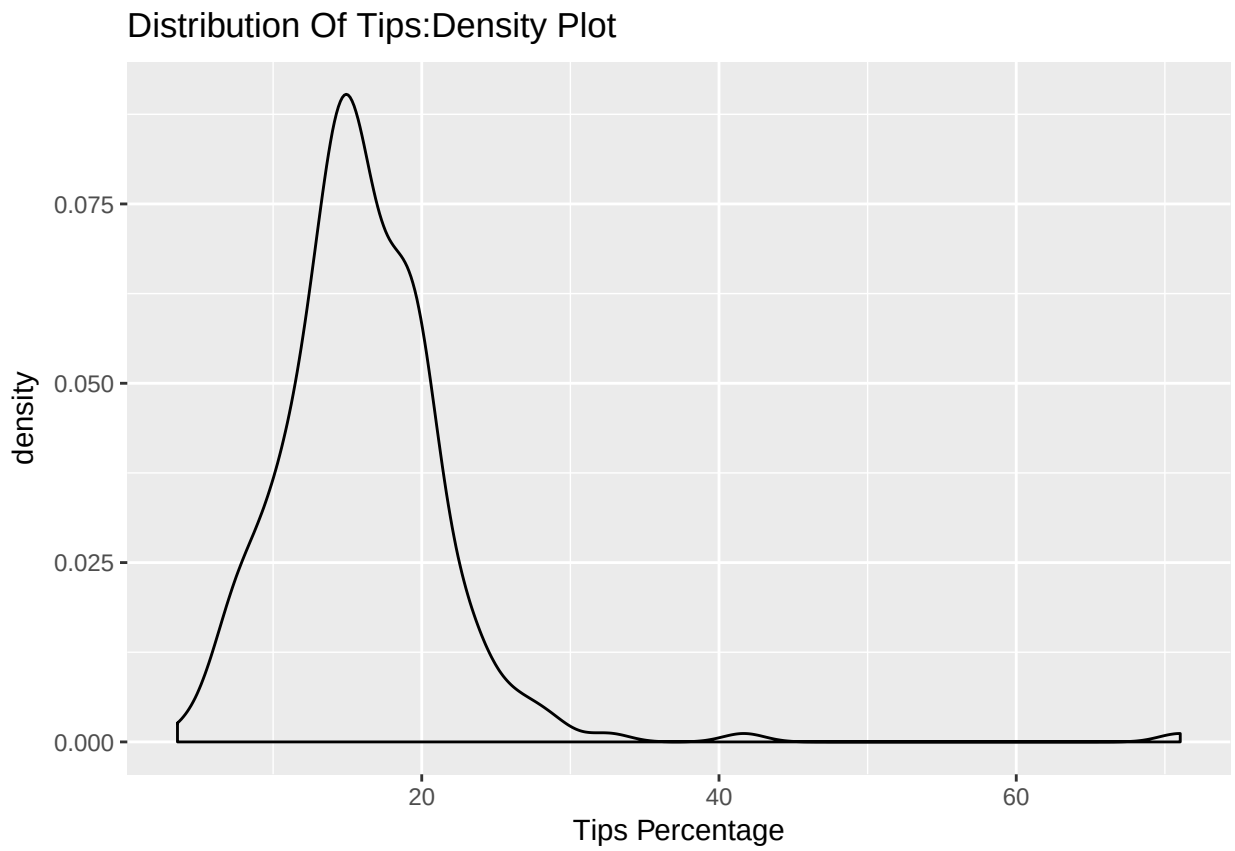
```
## Warning: package 'ggplot2' was built under R version 3.4.3
```

```
my_ggplot = ggplot(data = data,aes(x = tips_percent))
```

Center,spread and shape of Distribution

Creating Density plot

```
my_ggplot + geom_density(adjust=1) + xlab("Tips Percentage") +
ggtitle("Distribution Of Tips:Density Plot")
```



Observations

The mean of distribution is between 15 to 18. The distribution is right skewed and ranges from 2 to 70 appromximately.

We can verify this using summary statistics.

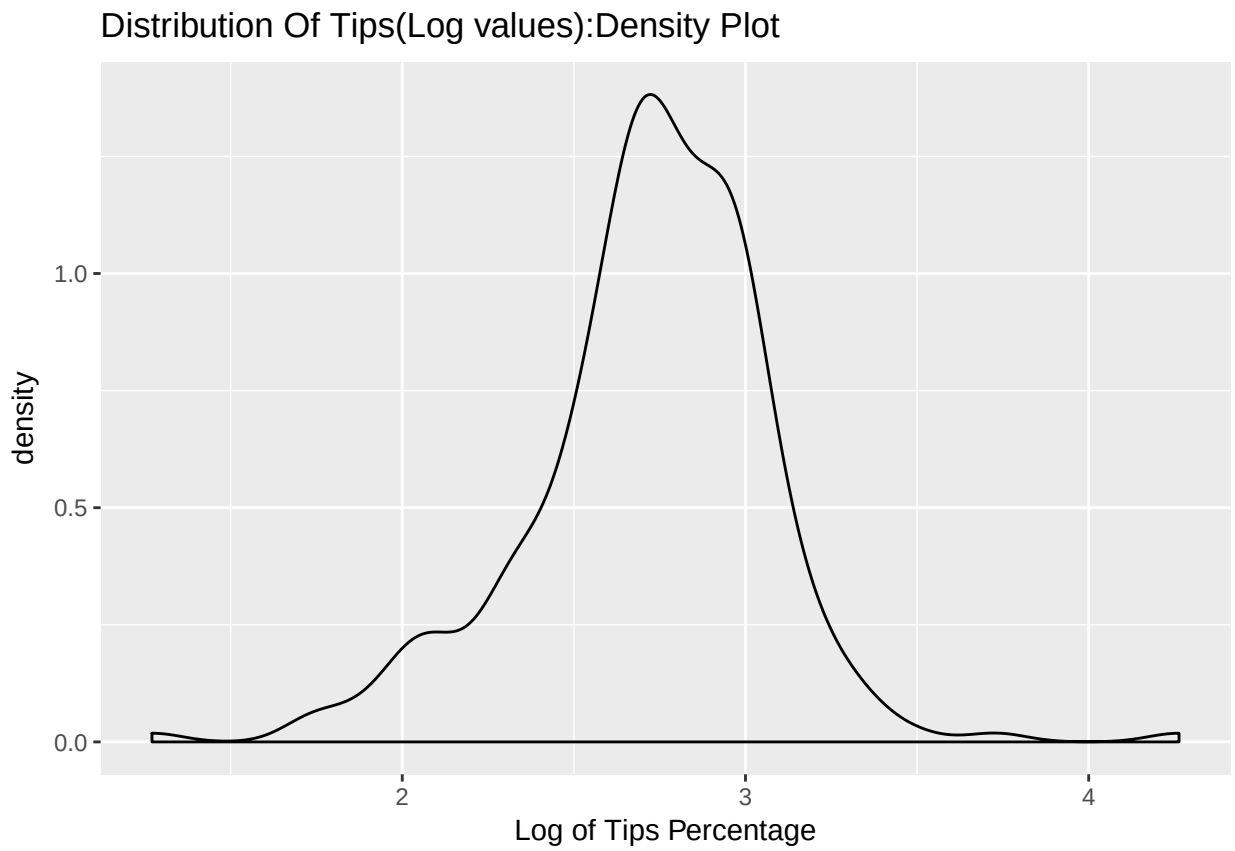
```
summary(data$tips_percent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.564  12.913  15.477  16.080  19.148  71.034
```

Performing Transformations

Since, the density Plot is right skewed. Therefore, we can perform log transformation to check if we can obtain a better symmetrical distribution.

```
my_log_ggplot = ggplot(data = data,aes(x=log(tips_percent)))
my_log_ggplot + geom_density(adjust=1) + xlab("Log of Tips Percentage") + ggtitle("Distribution Of Tips
```



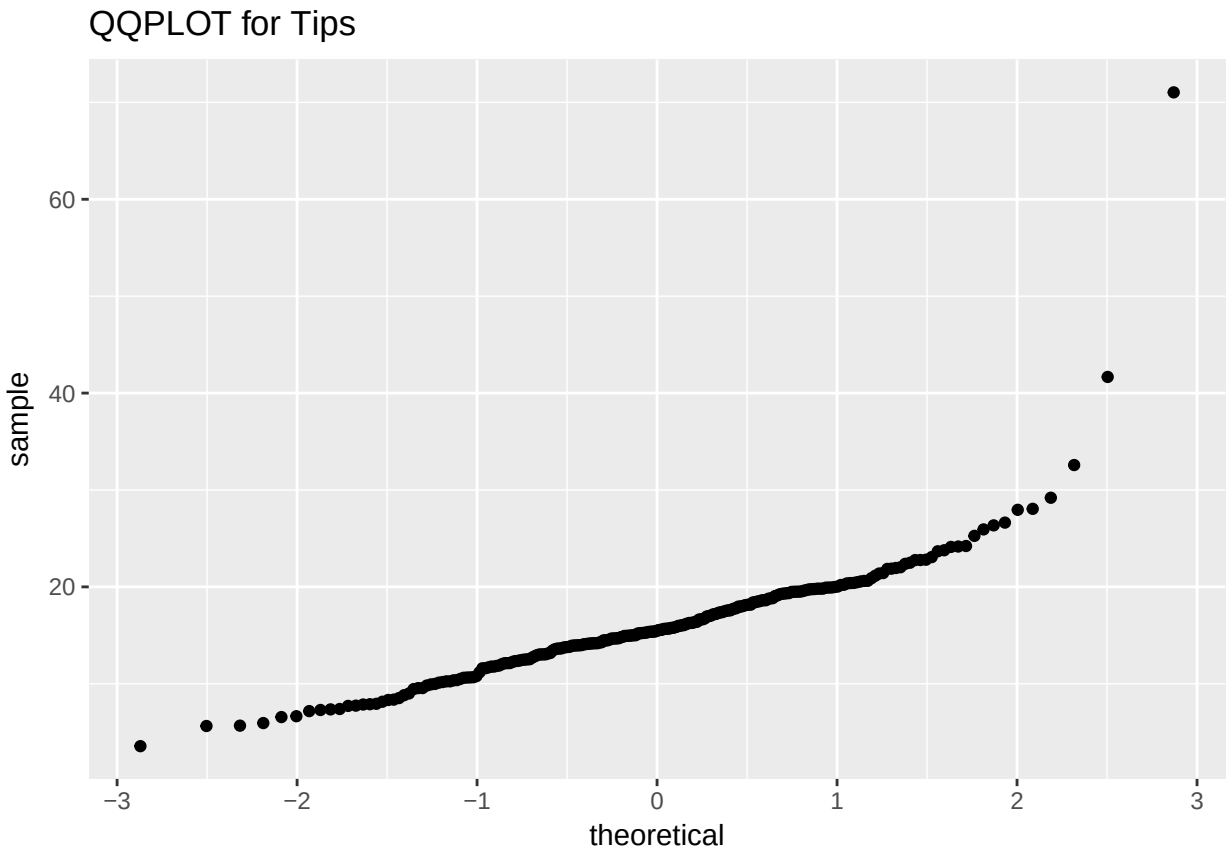
Observations

Taking log transformations give us a better symmetrical distribution.

Checking Normality of Data

QQPlot of data

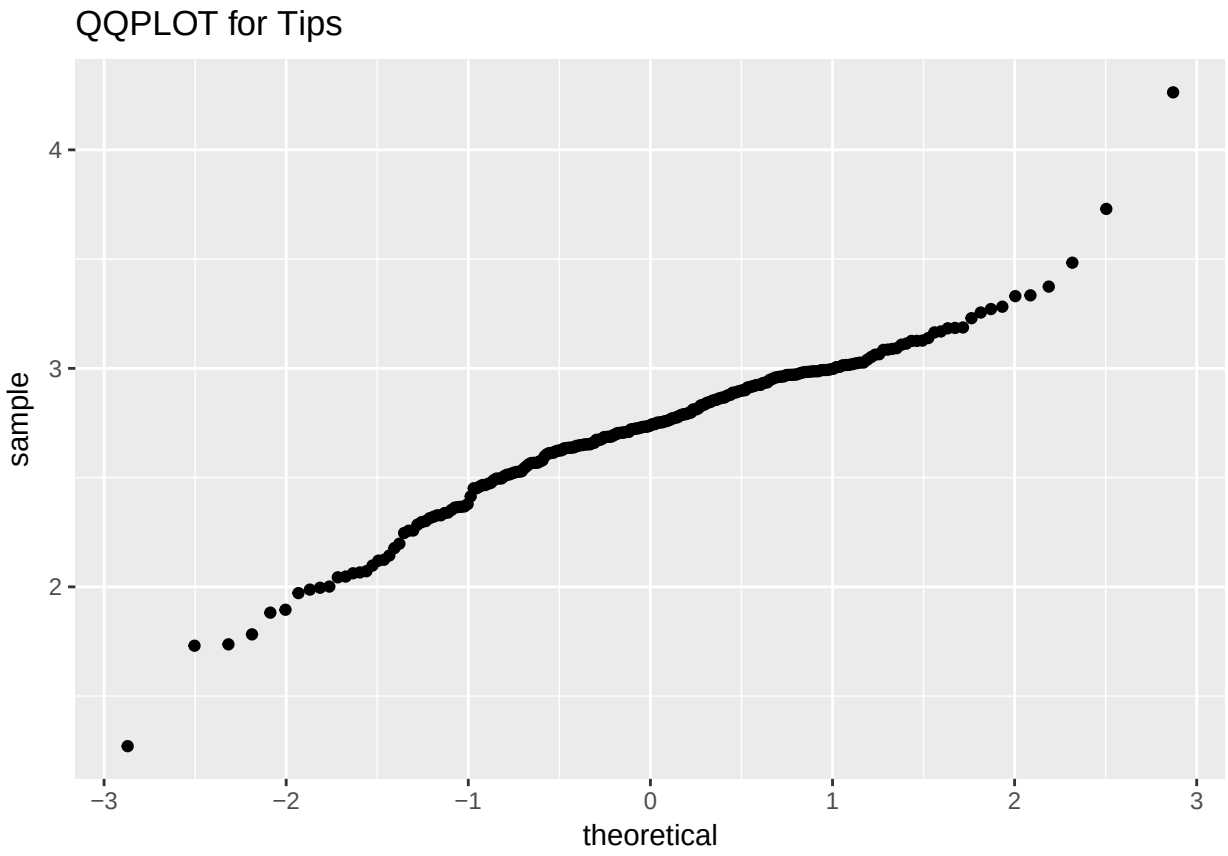
```
ggplot(data,aes(sample=tips_percent)) + stat_qq() + ggtitle("QQPLOT for Tips")
```



Checking Normality of Log Data

QQPlot of Log data

```
ggplot(data,aes(sample=log(tips_percent))) + stat_qq() + ggtitle("QQPLOT for Tips")
```



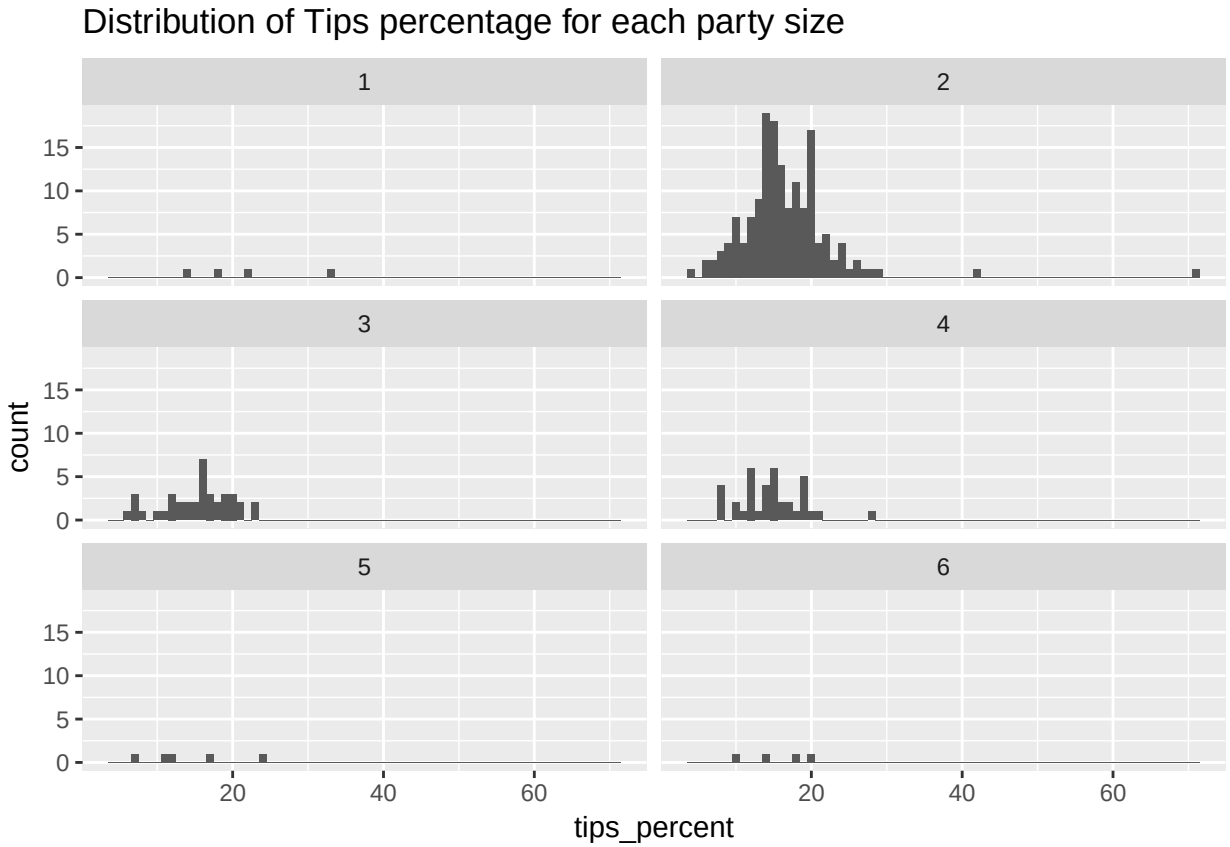
Observations

QQPlot of the transformed data has a better linearity than the original data.

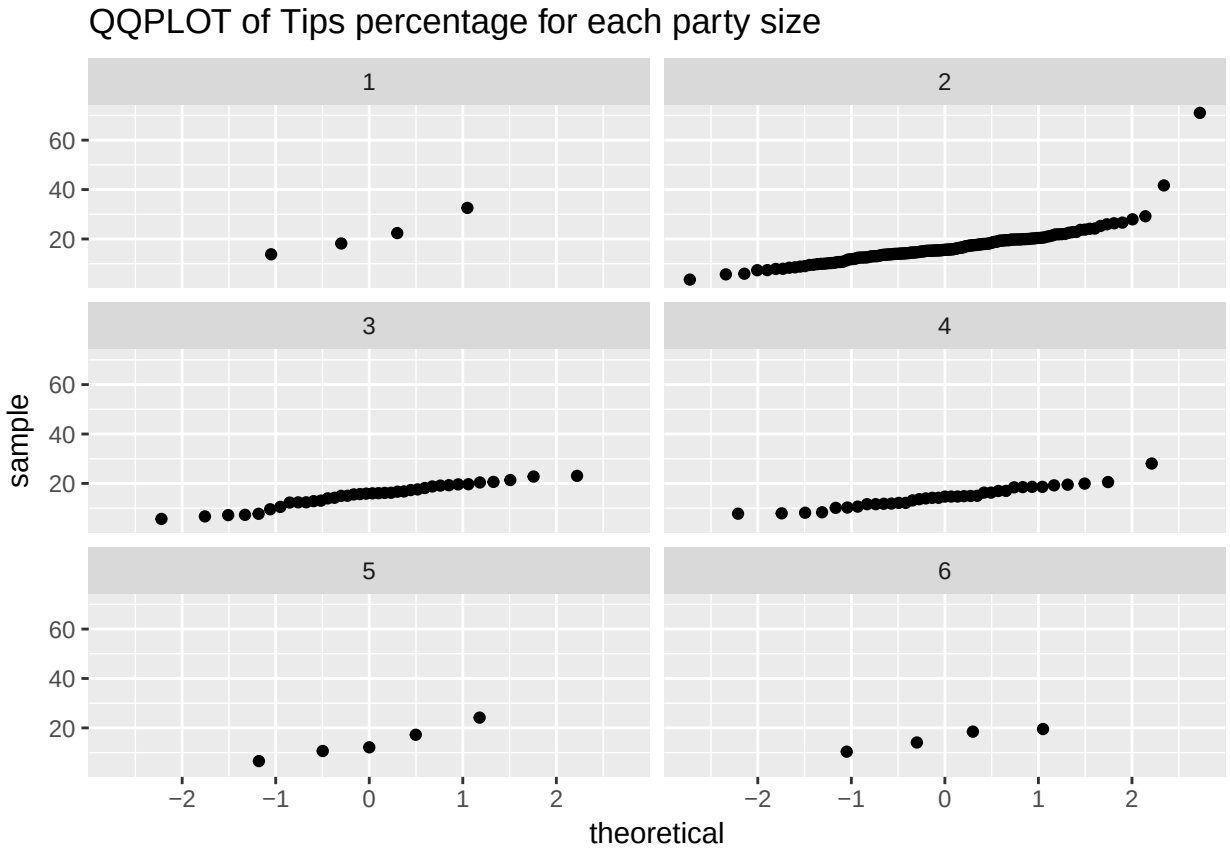
Question 2

Distribution of Percentage tipped for each Party Size.

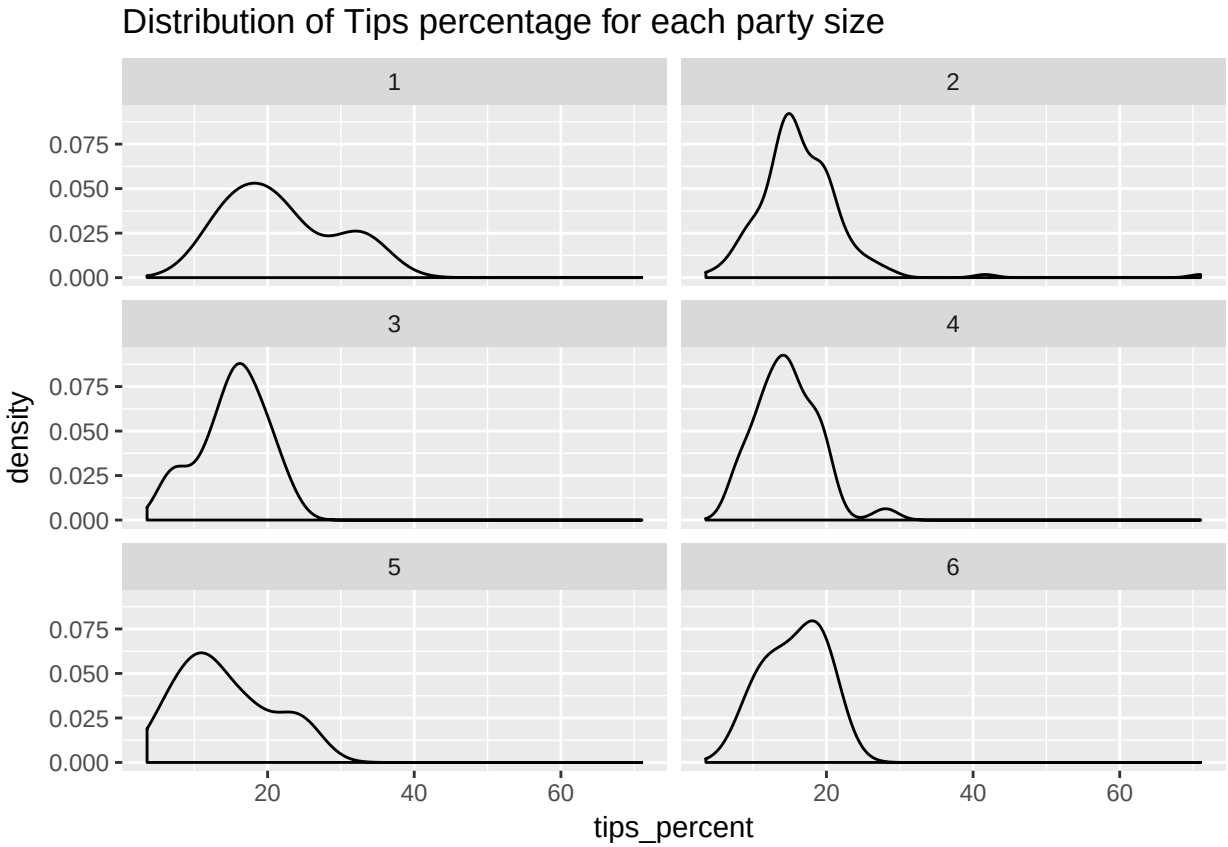
```
my_ggplot + geom_histogram(binwidth = 1) +
  ggtitle ("Distribution of Tips percentage for each party size") +
  facet_wrap(facets = ~data$size,ncol = 2)
```




```
ggplot(data,aes(sample=tips_percent))+
  stat_qq()+ggtitle("QQPLOT of Tips percentage for each party size")+
  facet_wrap(facets = ~data$size,ncol = 2)
```




```
my_ggplot + geom_density(adjust = 1) +
  ggtitle ("Distribution of Tips percentage for each party size") +
  facet_wrap(facets = ~data$size,ncol = 2)
```



Observations

From the density plot faceted graphs, we can observe that the distribution for 2,3 & 4 party sizes are similar.The density plots are right skewed with mean between 15-20. The QQplots also appear to be nearly straightish. However the distribution for other party sizes (1,5 and 6) are completely different than others. From the histogram faceted graphs, we can observe that the sample sizes of 1,5 & 6 party sizes are very small. This might be a reason for the variations in the distributions. Therefore we cannot make a strong conclusion about the distributions of party sizes 1,5 & 6. We need more data for them.

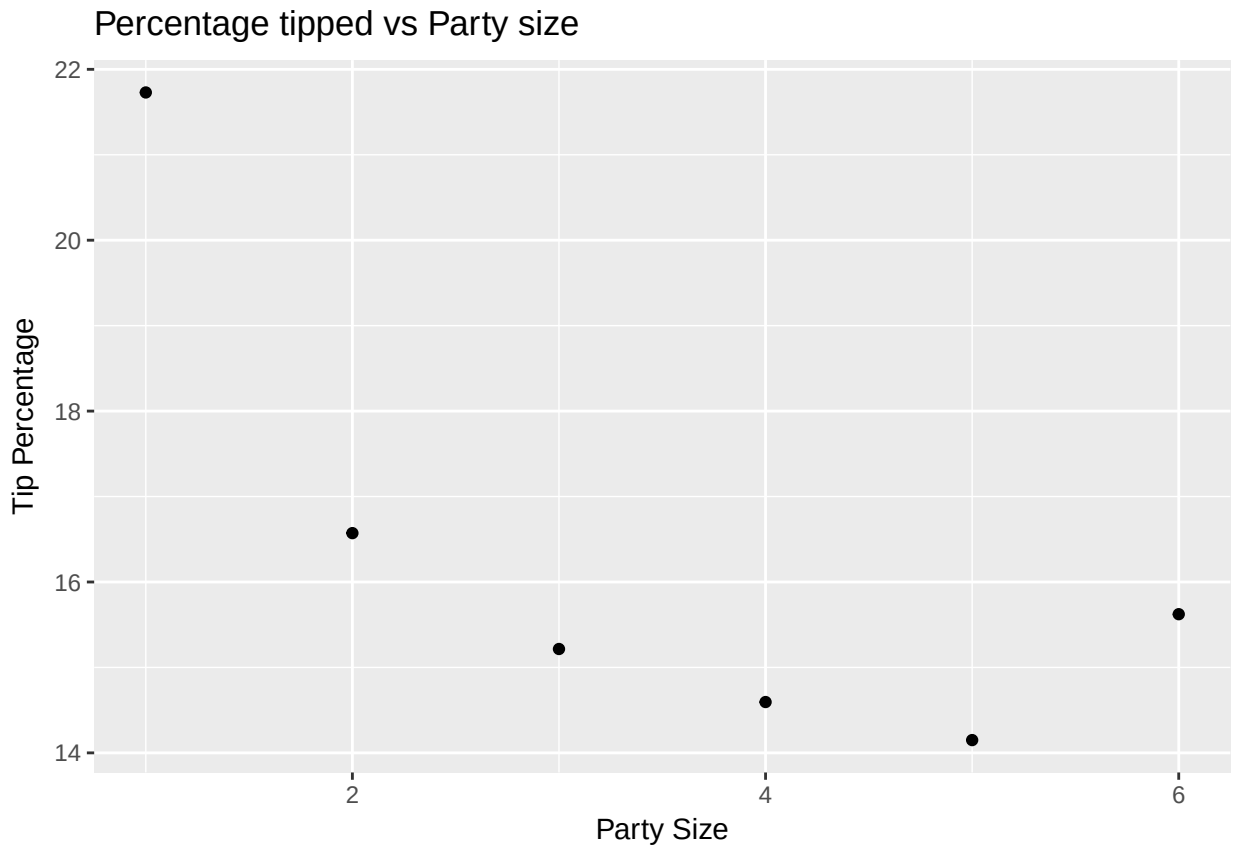
Question 3

Though our data is right skewed, but the data after transformation looks nearly symmetrical. Therefore mean is a good measure of center for the percentage tipped distributions.

```
size = data$size
tips.mean = aggregate(tips_percent ~ size,FUN = mean,data = data)
tips.mean$tips_percent
```

```
## [1] 21.72920 16.57192 15.21569 14.59490 14.14955 15.62292
```

```
ggplot(tips.mean,aes(x = tips.mean$size,y=tips.mean$tips_percent)) +
  geom_point() + xlab("Party Size") + ylab("Tip Percentage") +
  ggtitle("Percentage tipped vs Party size")
```



Observations

We can observe that there is a variation in central tendency for party sizes. As we have seen earlier, sample sizes for party sizes 1,5 & 6 are small. Hence this difference can be attributed as variation by chance. However the variations in party sizes 2,3,4 look to be real.