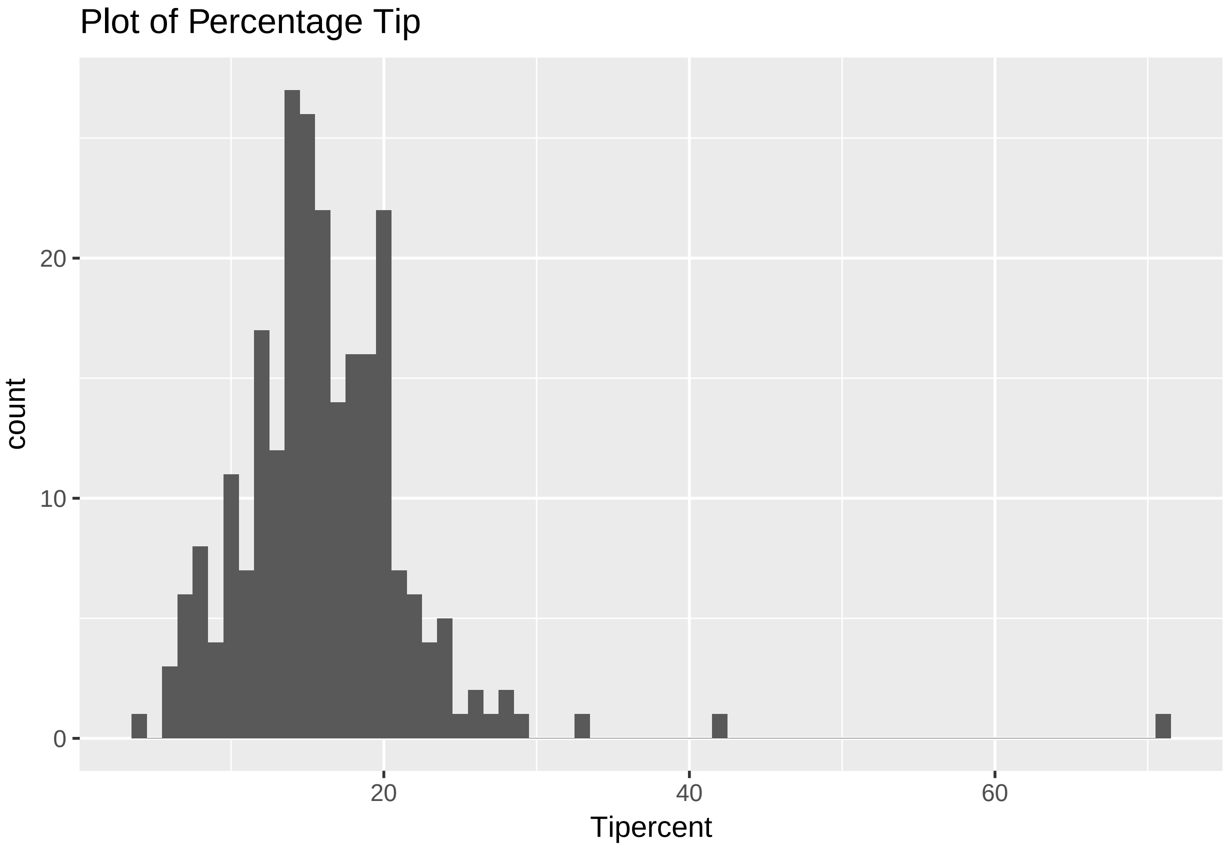
PS1 Solutions

Group members: Priyadarshini Vijjigiri, Saniya Ambavanekar

Q1

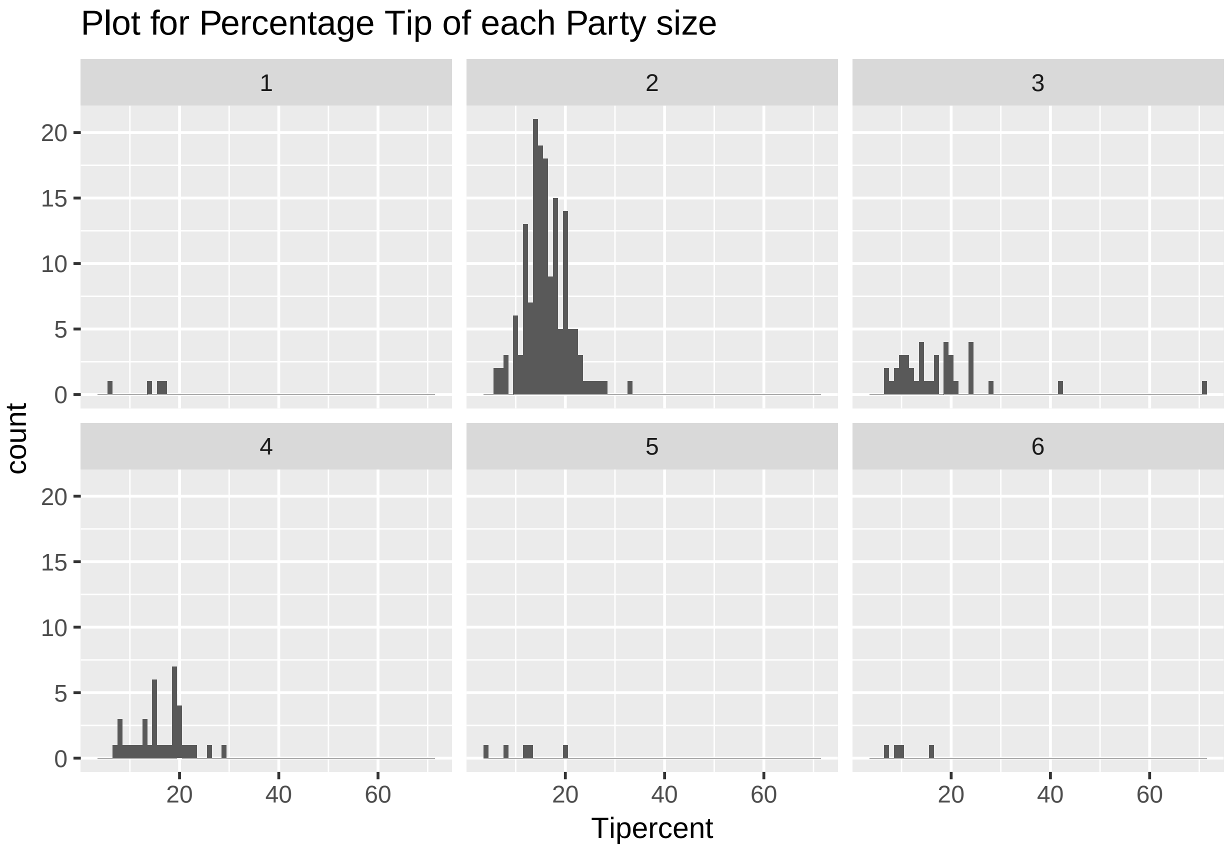
```
setwd("/Users/saniyaambavanekar/Downloads")
library(ggplot2)
dataframe = read.table("tips.txt", header = TRUE)
Tipercent = 100*(dataframe$tip)/(dataframe$total_bill)
ggplot(dataframe, aes(x = Tipercent)) + geom_histogram(binwidth = 1)+
ggtitle("Plot of Percentage Tip")
```



We can see from the graph that the centre is around 15, Spread is from 5 to 30 and it is visible that the distribution is not normal and is Right skewed and since the data is positive and rightly skewed we can apply log transformation to make it normal.

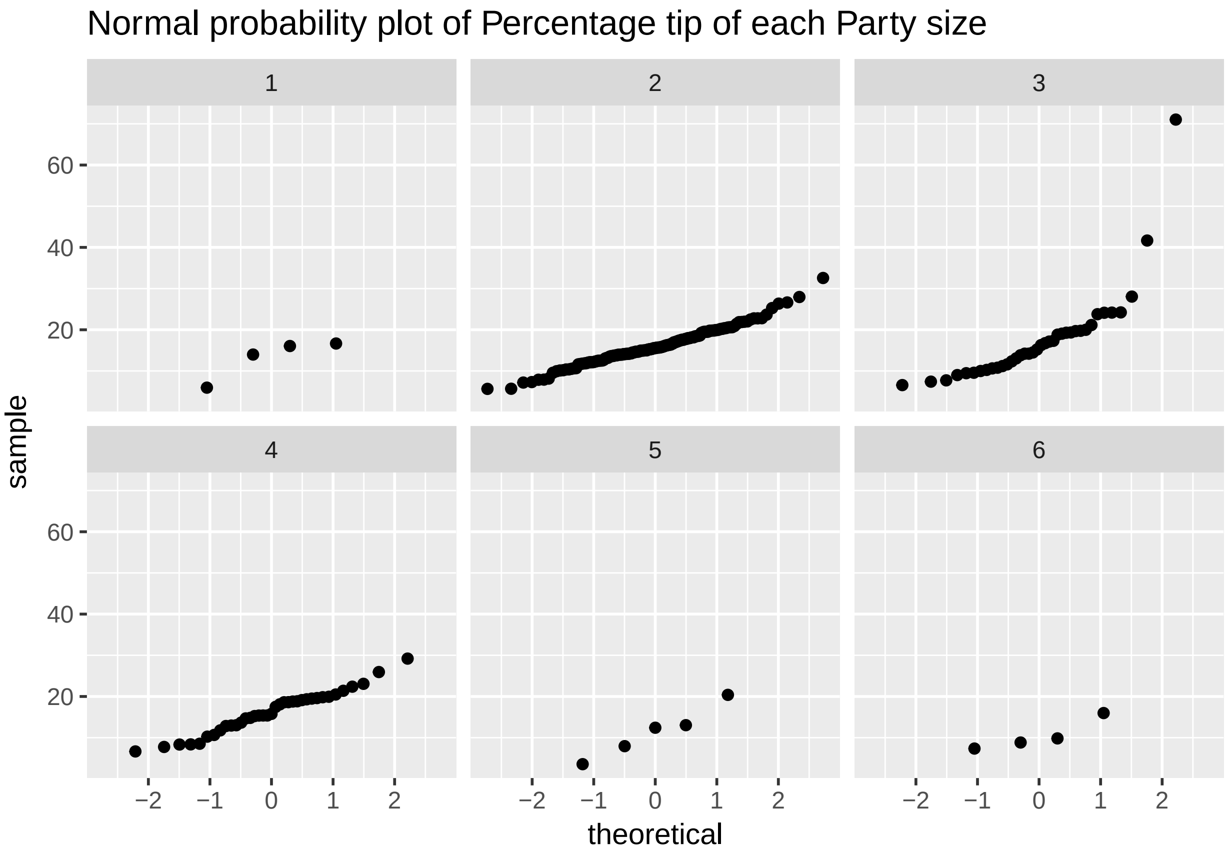
Q2

```
ggplot(dataframe, aes(x = Tipercent)) + geom_histogram(binwidth = 1) +
  facet_wrap(~size, ncol = 3)+ggtitle("Plot for Percentage Tip of each Party size")
```



Yes, there is difference between each distribution. We can observe that sample size for party size 2 is more and for party size 3(with some outliers in party size 3) and 4 sample size looks similar (with slight difference). Similarly sample size for party size 1 and 5 looks similar.

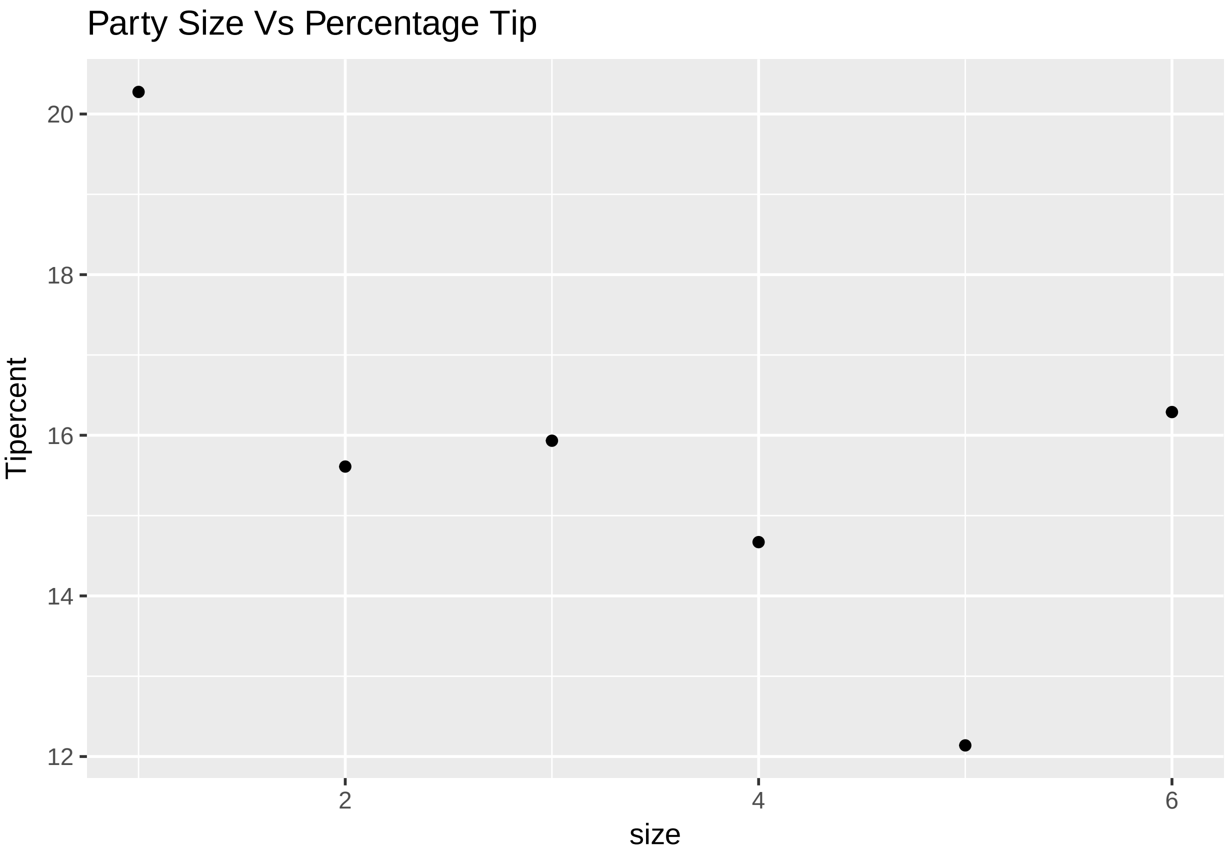
```
ggplot(dataframe,aes(sample=Tipercent))+stat_qq()+facet_wrap(~size,ncol=3)+
  ggtitle("Normal probability plot of Percentage tip of each Party size")
```



Q3

Since the distribtuions are not symmetric and has outliers, I have choosen median as a center for the percentage tipped distributions.

```
dataframe.median = aggregate(Tipercent ~ size, FUN = median, data = dataframe)
ggplot(dataframe.median, aes(x = size, y = Tipercent)) + geom_point()+
  ggtitle("Party Size Vs Percentage Tip")
```



The differences in centers for size 1 and 5 are far from 2,3,4,6

Which differences in the centers for different party sizes look real, and which can be reasonably explained by chance variation?

```
sample_size=as.data.frame(table(dataframe$size))
#  Var1 Freq
#1    1    4
#2    2  156
#3    3   38
#4    4   37
#5    5    5
#6    6    4
```

We cannot interpret real centers from different party size since each party size has different sample size. For party size 1,5,6 the sample size is very small from which we cannot interpret the actual increase or decrease in the percentage tip from the sample.The sample size of party size 2 is very large as compared to the party size 3 and 4. For the party size 3 and 4 though there is very less difference between their sample size , they showcase real scenario that as the party size increases there is decrease in the percentage tip as shown by the chart. Hence we need significant sample size for each party size to interpret real centers.