Q1

In [ ]:
```
df = read.table("tips.txt",header = T)
df$tippercentage = 100*df$tip/df$total_bill
library(ggplot2)
```

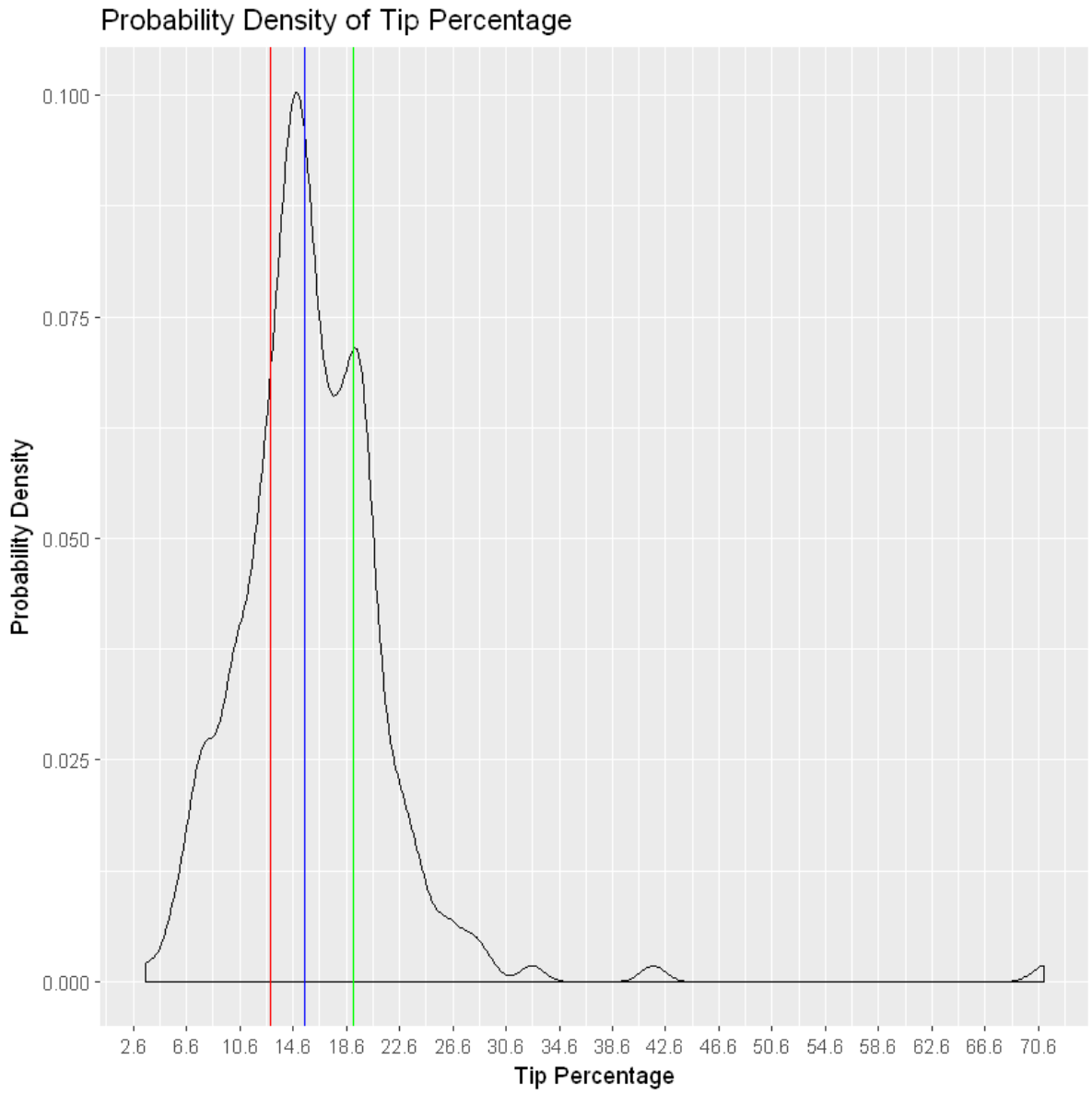
In [2]:
```
head(df)
```

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

In [3]:
```
summary(df$tippercentage)
```

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

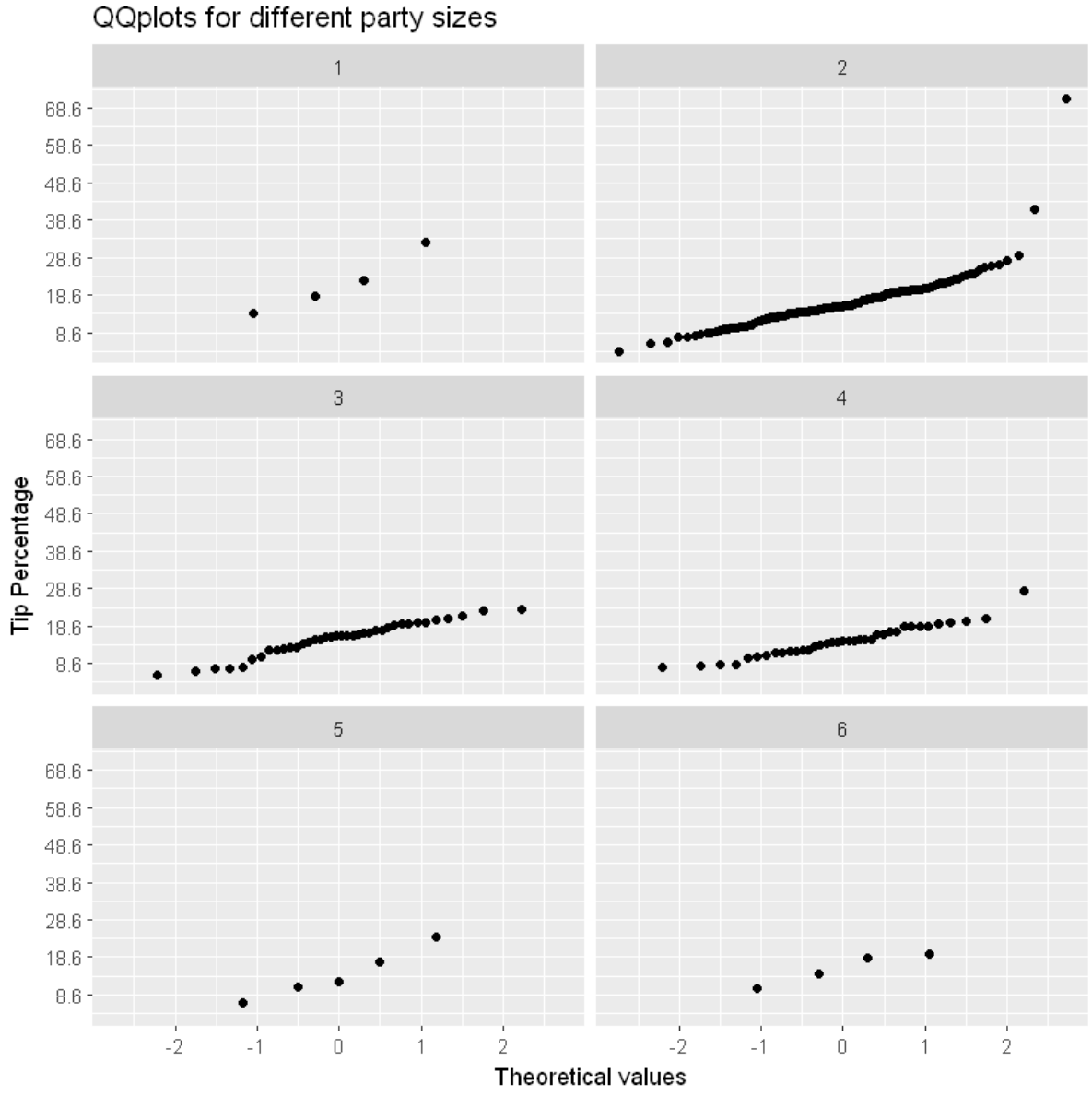
In [4]:
```
ggplot(df,aes(x=tippercentage)) + geom_density(adjust=0.65) +
  geom_vline(xintercept= quantile(df$tippercentage,probs=c(0.25,0.5,0.75)),col
or=c("red","blue","green"))+
  scale_x_continuous(breaks = round(seq(min(df$tippercentage)-5, max(df$tipper
centage)+5, by = 4),1)) +
  ylab("Probability Density") + xlab("Tip Percentage") + ggtitle("Probability
 Density of Tip Percentage")
```



- By observing the above plot it appears that the Tip Percentage is not normally distributed and is Right skewed.
- Since the data is right skewed mean is not an appropriate parameter to describe the center or spread of data.
- Redline – Q1 – 12.913
- Blueline – Median – 15.477
- Greenline – Q3 – 19.148
- We can clearly observe that (Median-Q1) ≠ (Q3-Median). Hence the data is not symmetric.

Q2

In [5]:
```
ggplot(df, aes(sample = tippercentage)) + stat_qq() + facet_wrap(~size, ncol =
 2)+
  scale_y_continuous(breaks = round(seq(min(df$tippercentage)-5, max(df$tipper
centage)+5, by = 10),1)) +
  xlab("Theoretical values") + ylab("Tip Percentage") + ggtitle("QQplots for d
ifferent party sizes")
```

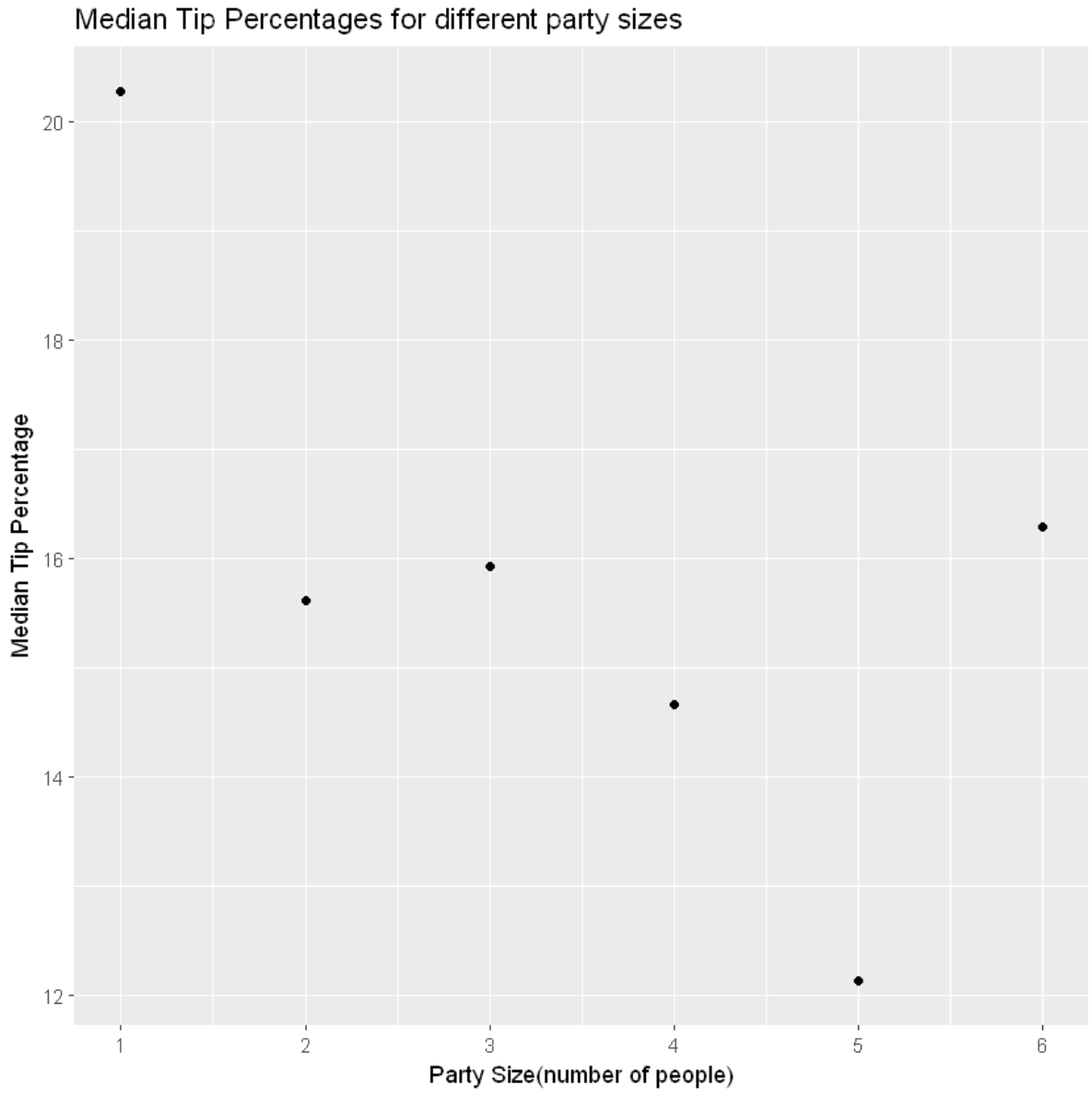


- There is clearly not enough data for party sizes = 1,5 & 6 to make any interpretations.
- The distributions for the party sizes = 2,3 & 4 appear similar with their means centered around 15.

Q3

In [6]:
```
df.medians = aggregate(tippercentage ~ size, FUN = "median", data = df)

ggplot(df.medians,aes(x=size,y=tippercentage))+geom_point()+
  scale_x_continuous(breaks=seq(0,7,1))+
  xlab("Party Size(number of people)")+
  ylab("Median Tip Percentage")+
  ggtitle("Median Tip Percentages for different party sizes")
```



• Since the data is not symmetric and with significant outliers Median is appropriate choice to describe the measure of center.

• For the party sizes = 2,3 & 4 the centers (medians) are around 15 and look real.

• For party size = 6 the center (median) is around 15 but it is only because of chance variation (do not have enough sample size)

• For party sizes = 1 & 5 the centers (medians) are extreme values (highest and lowest) but the sample sizes are not large enough to make any conclusions.