Problem Set 1

Shailendra Patil

1/12/2018

1

Loading the data

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
 mydata = read.table("/Users/shailendrapatil/Spring2018/EDA/PS1/tips.txt",header=TRUE)
```

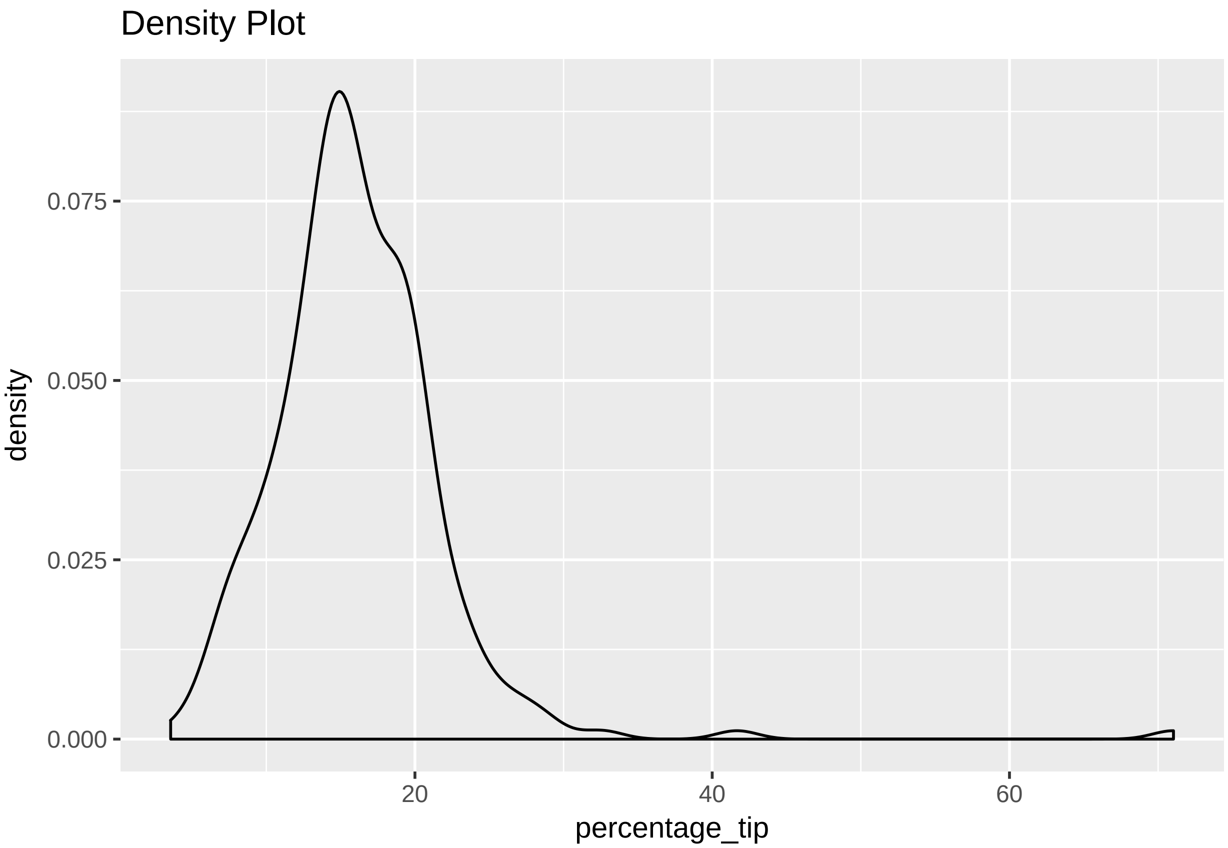
Adding the percentage tip column

```
percentage_tip=(mydata$tip/mydata$total_bill)*100
mydata<-cbind(mydata, percentage_tip)
head(mydata)
```

```
##   total_bill  tip    sex smoker day   time size percentage_tip
## 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
```

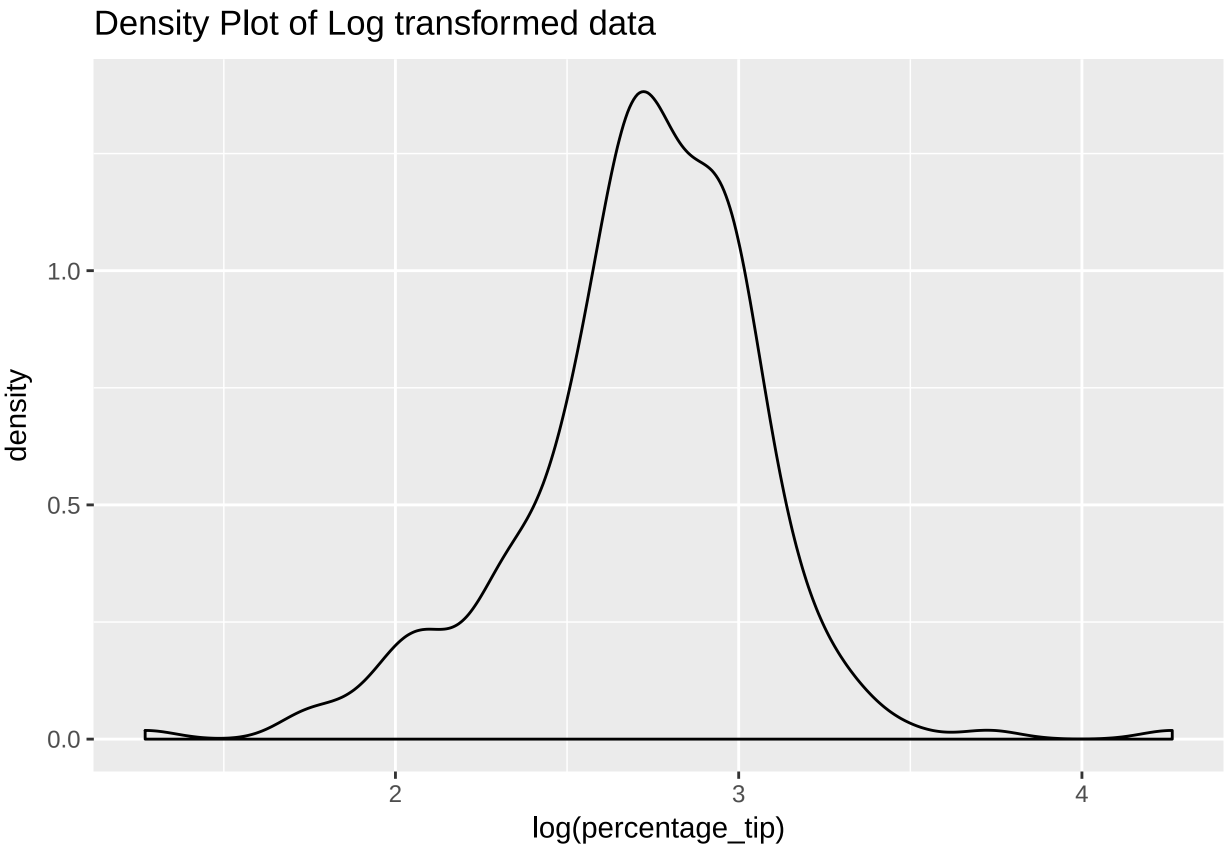
Drawing the density plot using adjust as 1 to make the plot smooth

```
ggplot(mydata, aes(x = percentage_tip)) + geom_density(adjust = 1)+ggtitle("Density Plot")
```



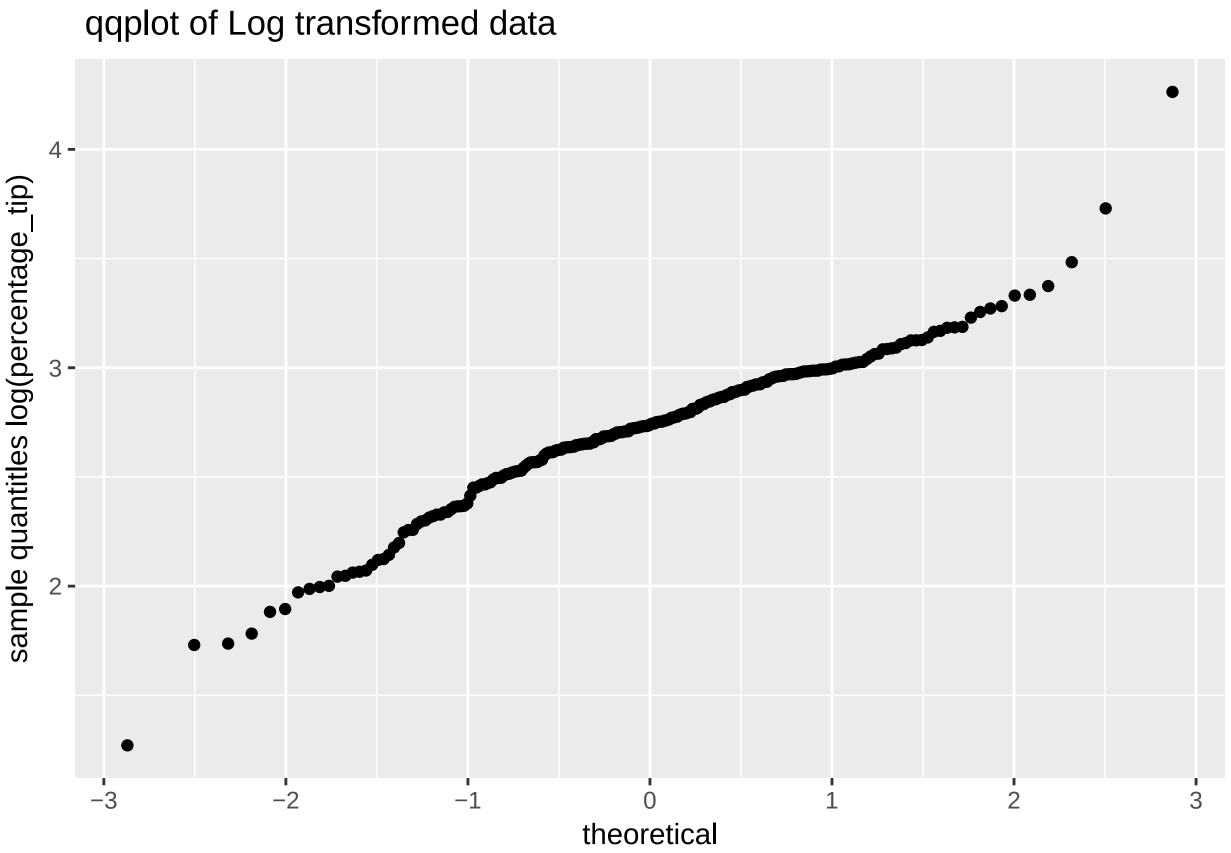
The above graph looks **right skewed**. Let us check if the **log transformation** of the data is normal or not.

```
ggplot(mydata, aes(x = log(percentage_tip))) +
  geom_density(adjust = 1)+
  ggtitle("Density Plot of Log transformed data")
```



The log transformation looks **nearly norma**l. But the **actual percentage tip is not a normal distribution**. Let us check the qqplot for the log transformation of perecentage tipped

```
ggplot(mydata, aes(sample = log(percentage_tip)))+
stat_qq()+ylab("sample quantitles log(percentage_tip)")+
ggtitle(" qqplot of Log transformed data")
```



The qqplot of the log transformation is **nearly linear** and hence we can say the **transformed data is nearly normal**.

The density plot shows the percentage tip is right skewed and the log transformation is nearly normal. And hence we can give the center(mean) of the data. As seen from the density plot most of the values are between 3 and 25, and hence mean is some value between them. And we can see the spread of the data as well, the lowest percentage tip as seen from plot density is somewhere close to 3-5 and the highest percentage tip is close to 70. And the actual distribution is right skewed.

We can verify the above details using summary statistics

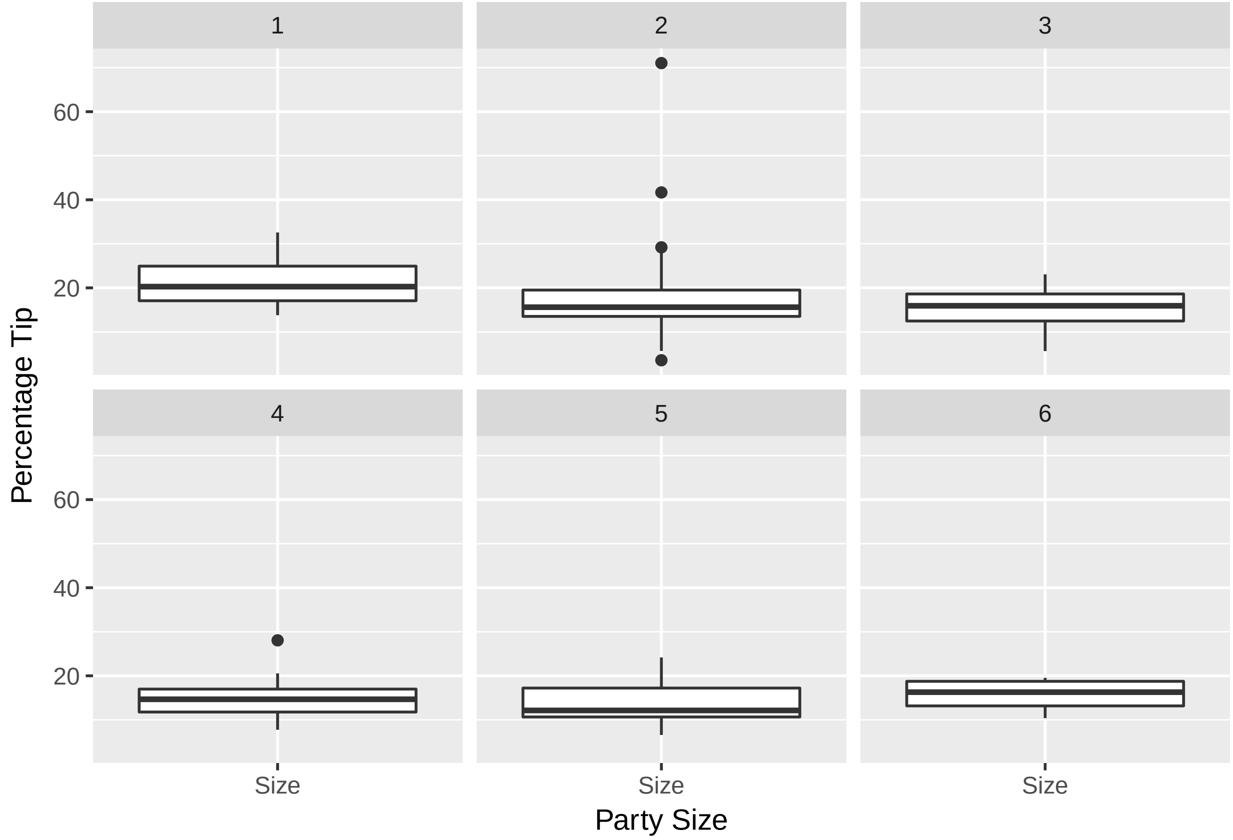
```
summary(mydata$percentage_tip)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.564  12.910  15.480  16.080  19.150  71.030
```

2

Let us check box plot using faceted plot for different party size:

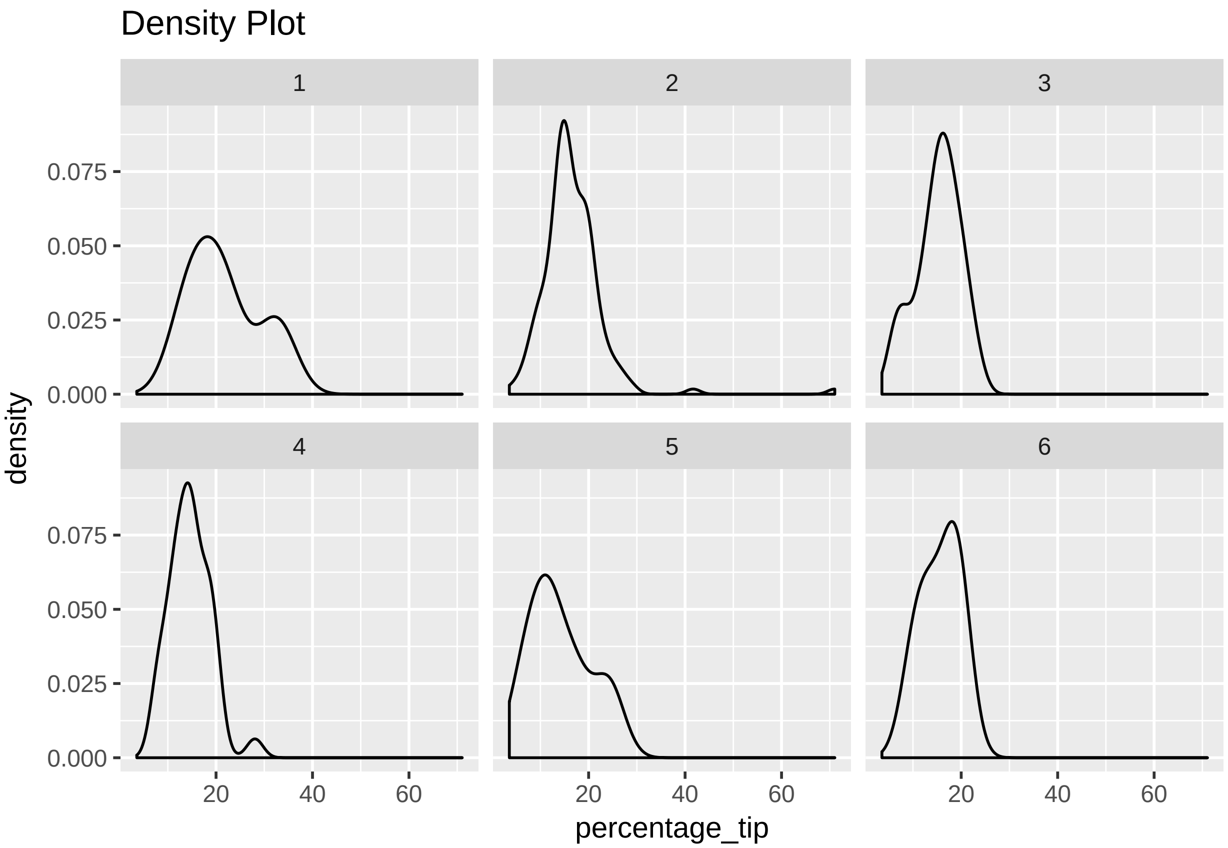
```
ggplot(mydata, aes(x = "Size", y = percentage_tip))+
geom_boxplot() +xlab("Party Size") +
ylab("Percentage Tip")+facet_wrap(~size,ncol = 3)
```



From the box plots, we can see that, a single person tends to tip higher compared to others. Party size 2, 3 and 4 have almost similar percentage tip. For party size 5 we can see that median value is close to 10 and majority values are above that. And for party size 6 median value is close to 20. So from box plot party size 2,3 and 4 seem to have similar distribution.

let us check the distribution using density plot and see if there is anything that we can conclude from those graphs

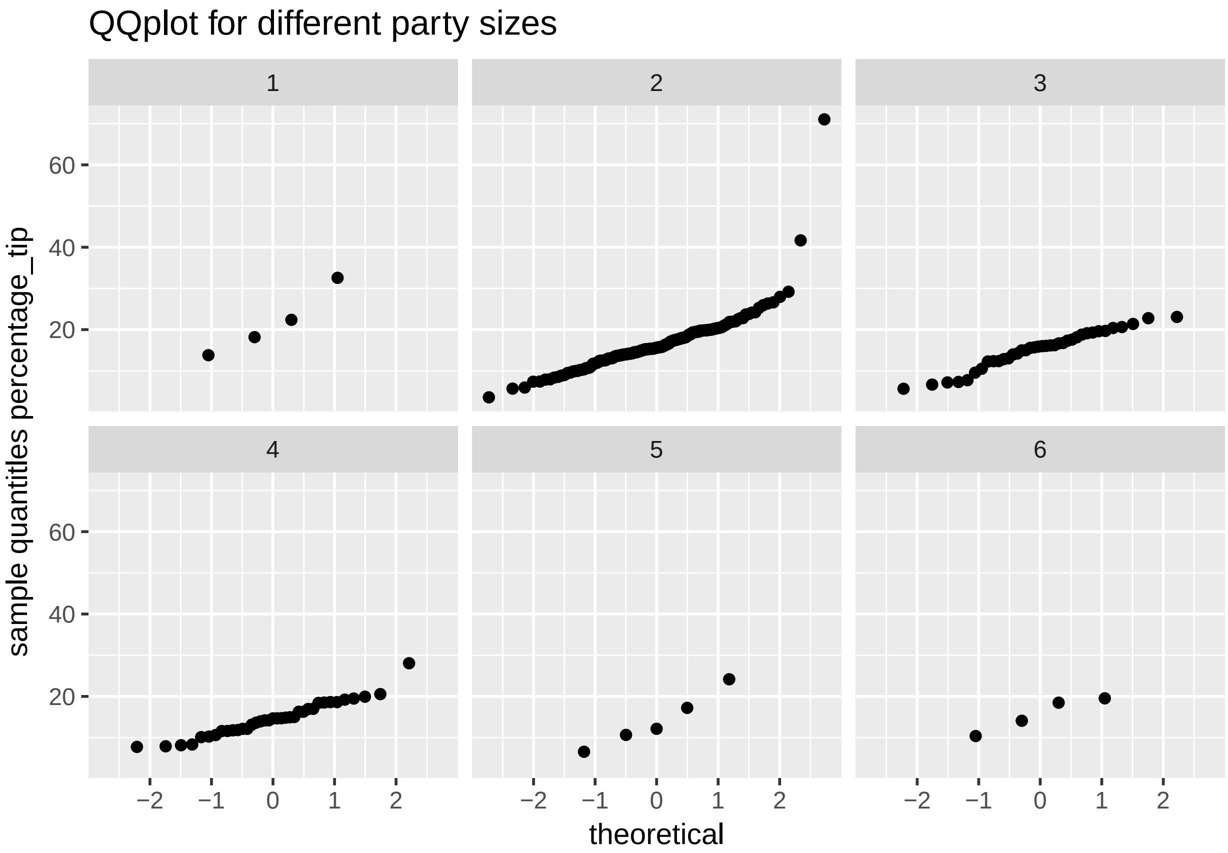
```
ggplot(mydata, aes(x = percentage_tip)) +
geom_density(adjust = 1)+
ggtitle("Density Plot")+facet_wrap(~size,ncol = 3)
```



The above density plot show that for party size 2,3 and 4 the distribution are right skewed , with peak somewhere in between 10 and 20, But for party size 1 , the plot doesn't exactly look like right skewed and for plot 5 the distribution is right skewed with its peak close to 10 and for party size 6 the distribution is right skewed with peak close to 20. From the density plot we can say that party size 2, 3 and 4 have similar distributions and party size 1,5 and 6 seem to have different distribution.

Let us check the qqplot for the same and check if we can come to any other conclusion on the data

```
ggplot(mydata, aes(sample = percentage_tip))+
stat_qq() + ylab("sample quantitles percentage_tip")+
ggtitle("QQplot for different party sizes")+
facet_wrap(~size,ncol = 3)
```

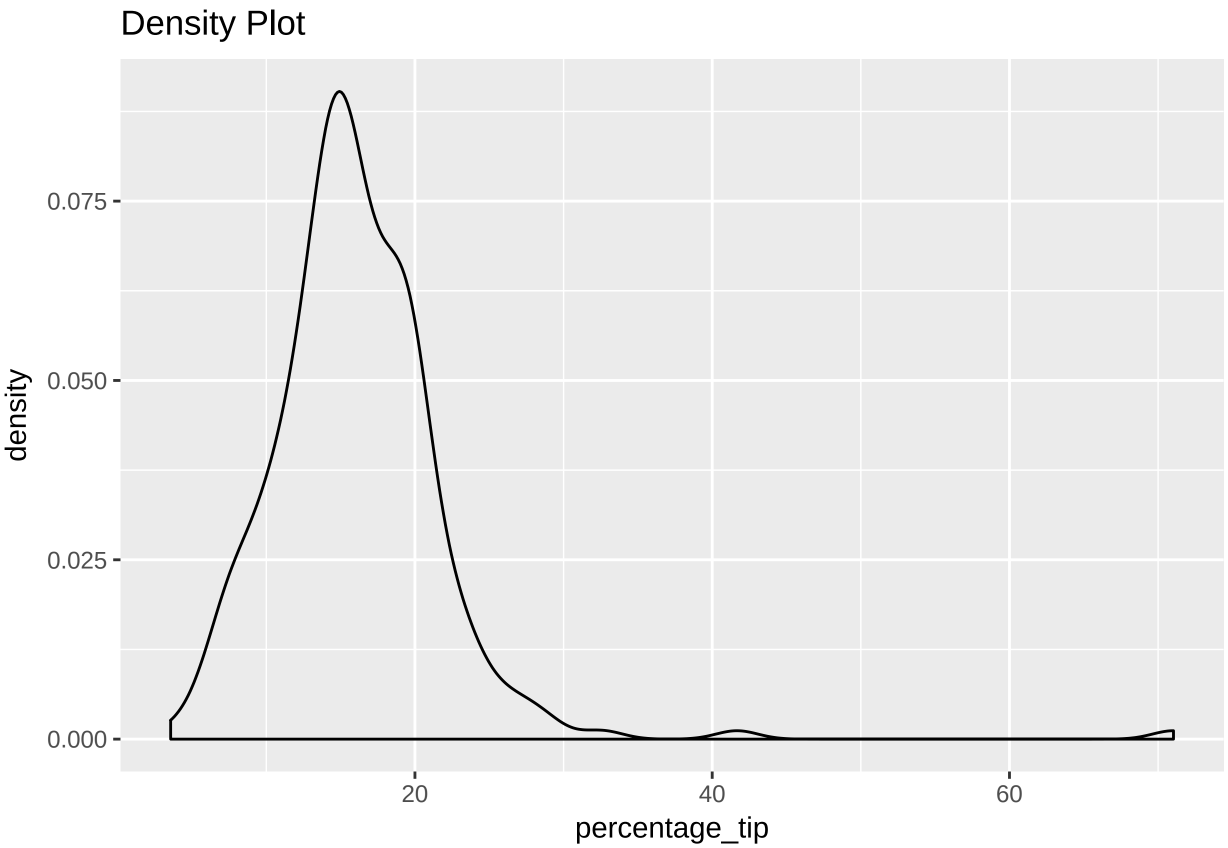


From the above graphs, we cannot conclude on all the analysis made so far for part size 1,5 and 6 as the sample size for these categories is very less. In order to comment or come to any kind of conclusion on a distribution, the sample size should atleast be comprable to that of other samples. In this case sample size doesnt even comprise 3% of the actual data.Hence it doesnt make any sense to comment on party size(1,5 and 6)

3

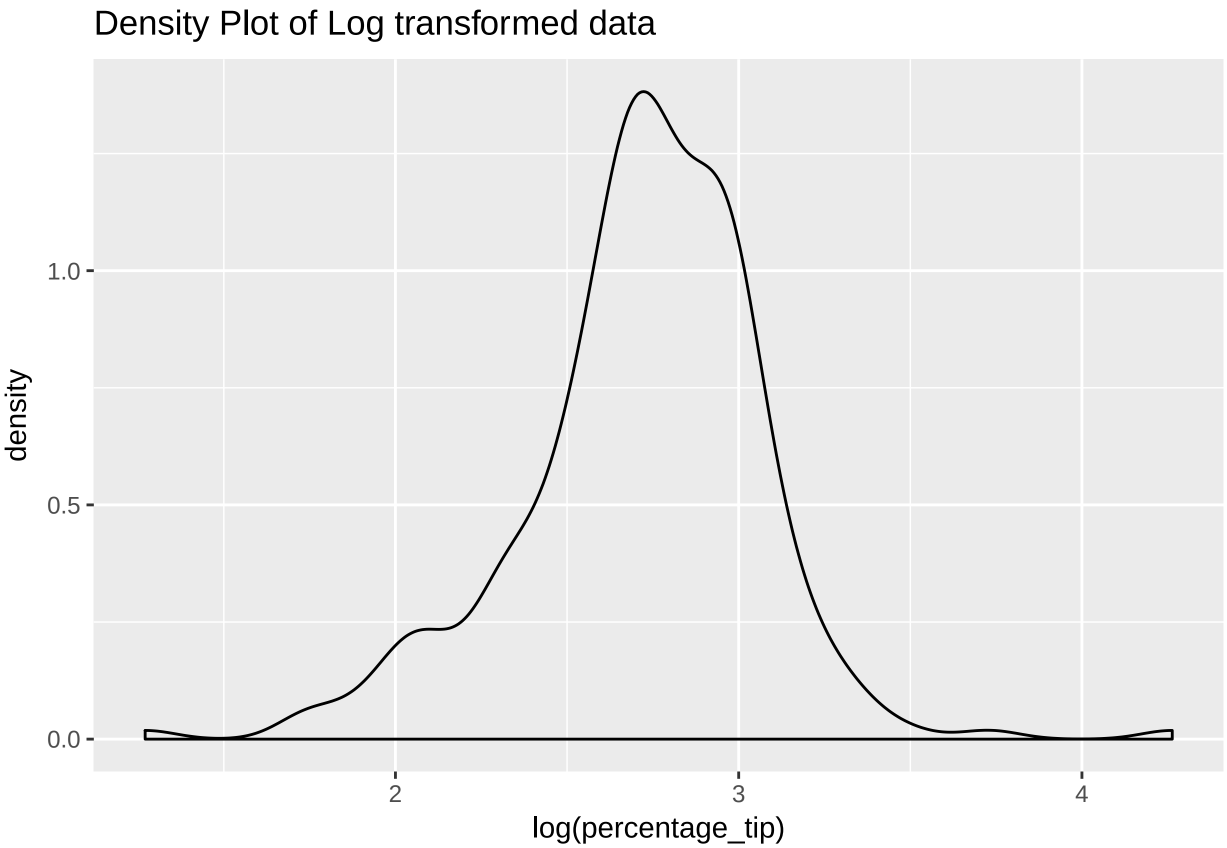
Drawing the density plot using adjust as 1 to make the plot smooth

```
ggplot(mydata, aes(x = percentage_tip)) + geom_density(adjust = 1)+ggtitle("Density Plot")
```



The above graph looks **right skewed**. Let us check if the **log transformation** of the data is normal or not.

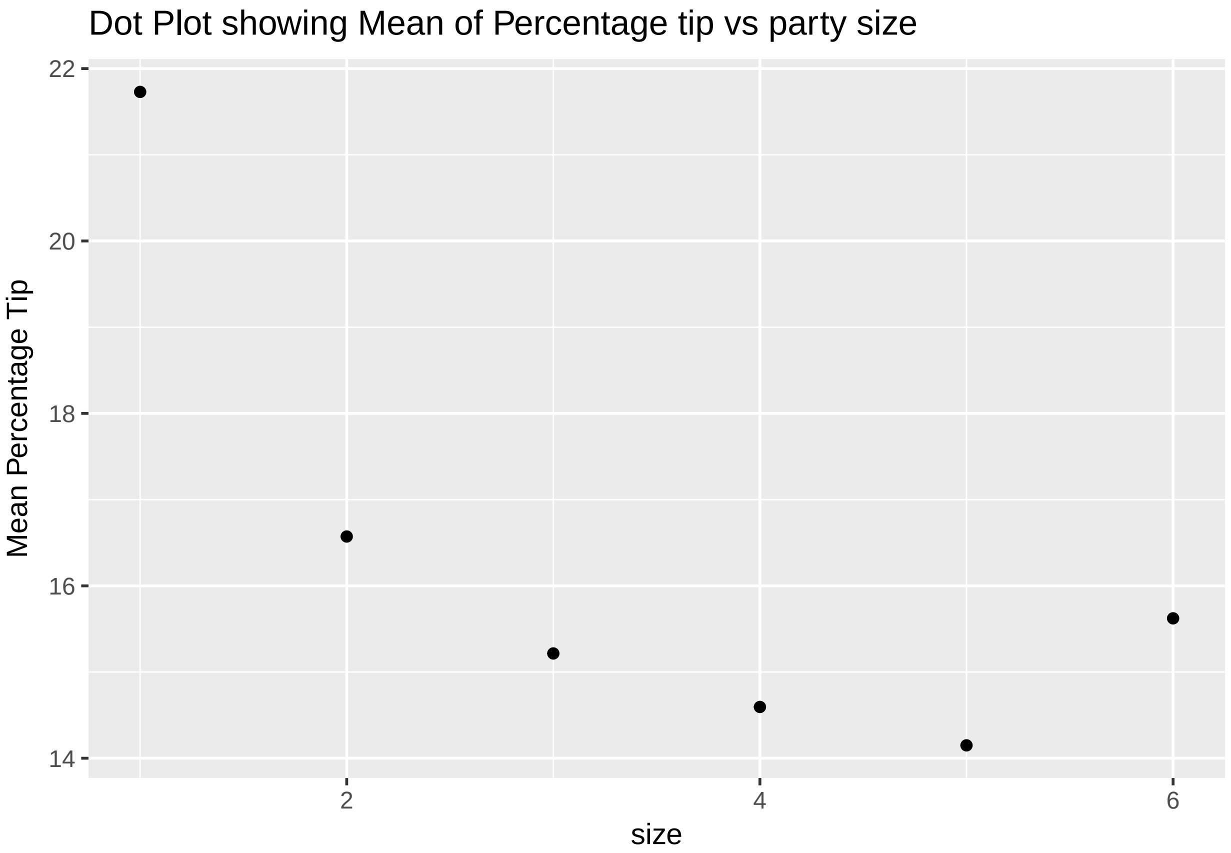
```
ggplot(mydata, aes(x = log(percentage_tip))) +
  geom_density(adjust = 1)+
  ggtitle("Density Plot of Log transformed data")
```


Density Plot of Log transformed data

The log transformation looks **nearly norma**l. But the **actual percentage tip is not a normal distribution**. And hence we choose **Mean** as a center for the distribution

Finding center using aggregate function

```
percentage_tip.means = aggregate(percentage_tip ~ size,
                                 FUN = mean, data = mydata)
ggplot(percentage_tip.means,
       aes(x = size, y = percentage_tip)) + geom_point()+ylab("Mean Percentage Tip")+
ggtitle("Dot Plot showing Mean of Percentage tip vs party size")
```



As seen earlier party size 1,5 and 6 sample size is too small and hence we can say that difference in mean percentage tip in these party size can be explained by chance variation and difference in mean percentage tip for party size 2,3 and 4 looks to be real.