use same file for this assignment which you used for practice assignment for Univariate Statistics

Does your data follow normal distribution?

Note: you can revisit this question once you have gone through hypothesis testing module

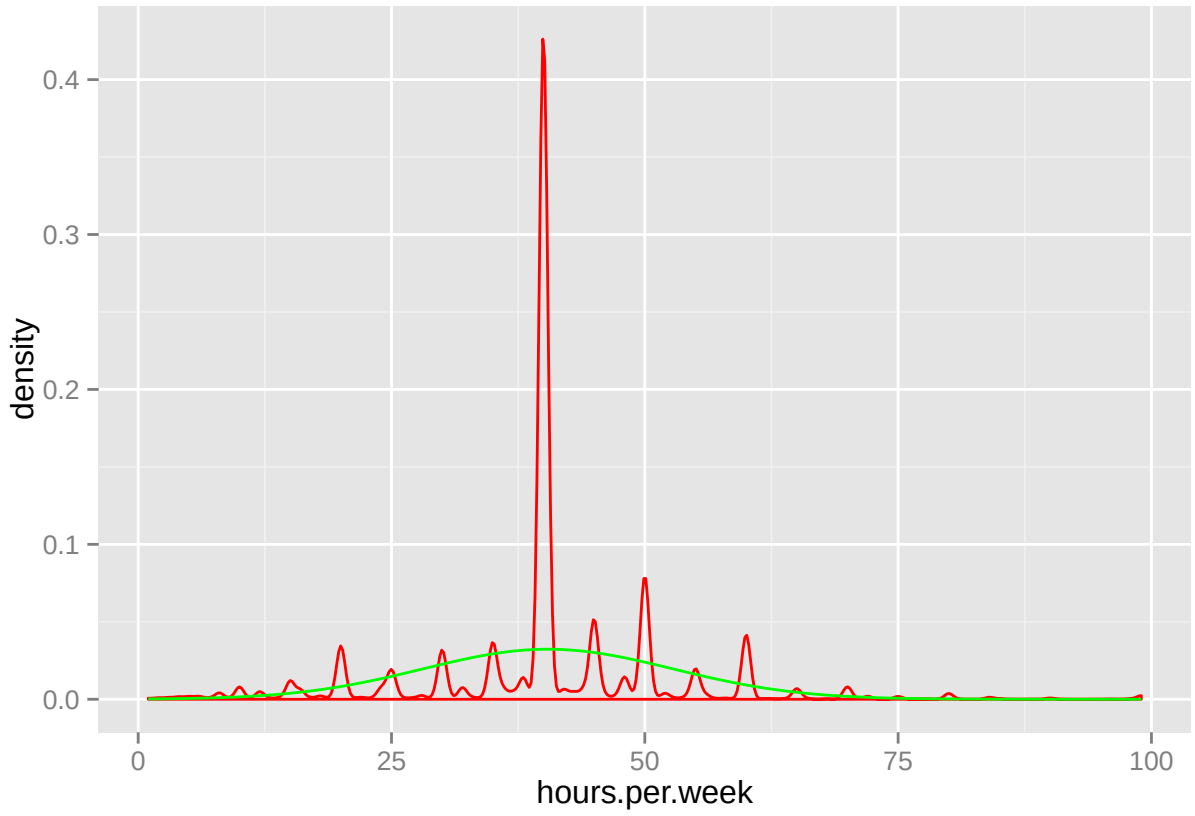
Find out if variable hours.per.week follows normal distribution using density plots. Your plot should like this:

```
setwd("/Users/lalitsachan/Desktop/March onwards/CBAP with R/Data/")
# You'll have to chose path accroding to location of file in your machine

d=read.csv("census_income.csv",stringsAsFactors = F)
library(ggplot2)

d$temp=dnorm(d$hours.per.week,mean=mean(d$hours.per.week),sd=sd(d$hours.per.week))

ggplot(d,aes(x=hours.per.week))+
  geom_density(color="red")+
  geom_line(aes(x=hours.per.week,y=temp),color="green")
```

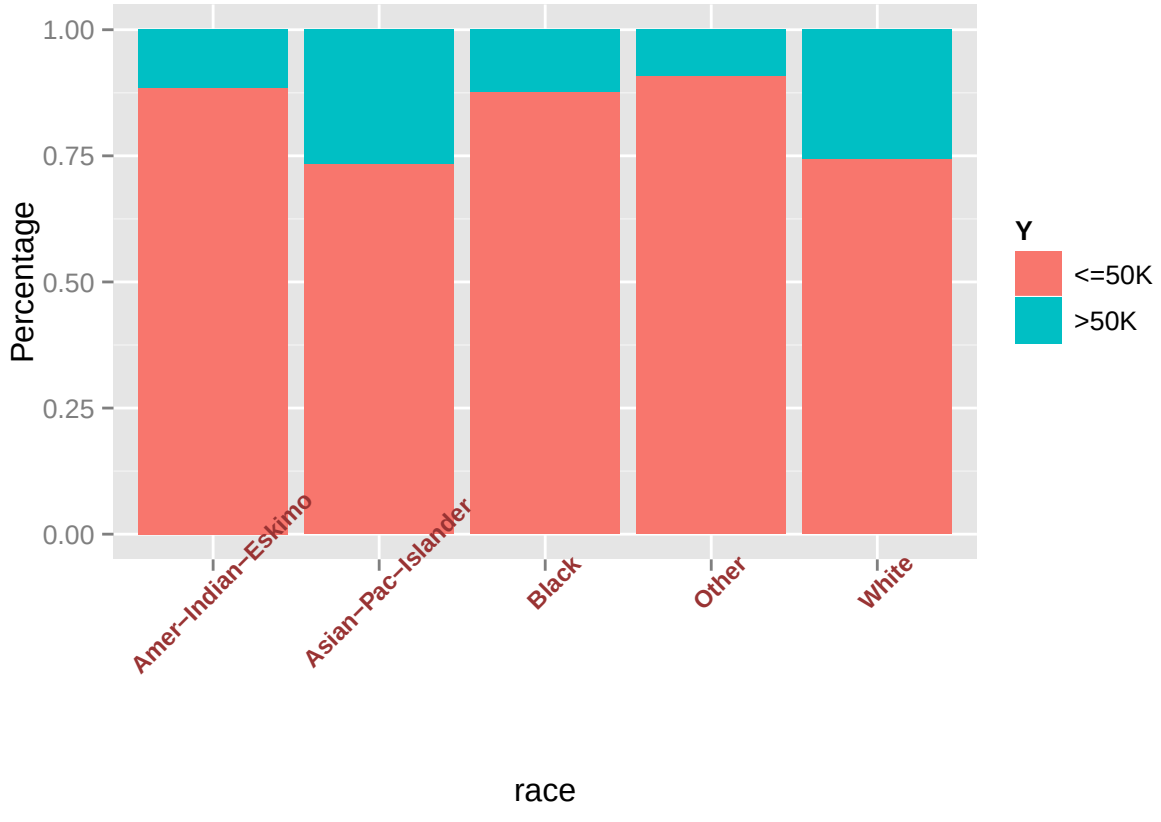



```
d$temp=NULL
```

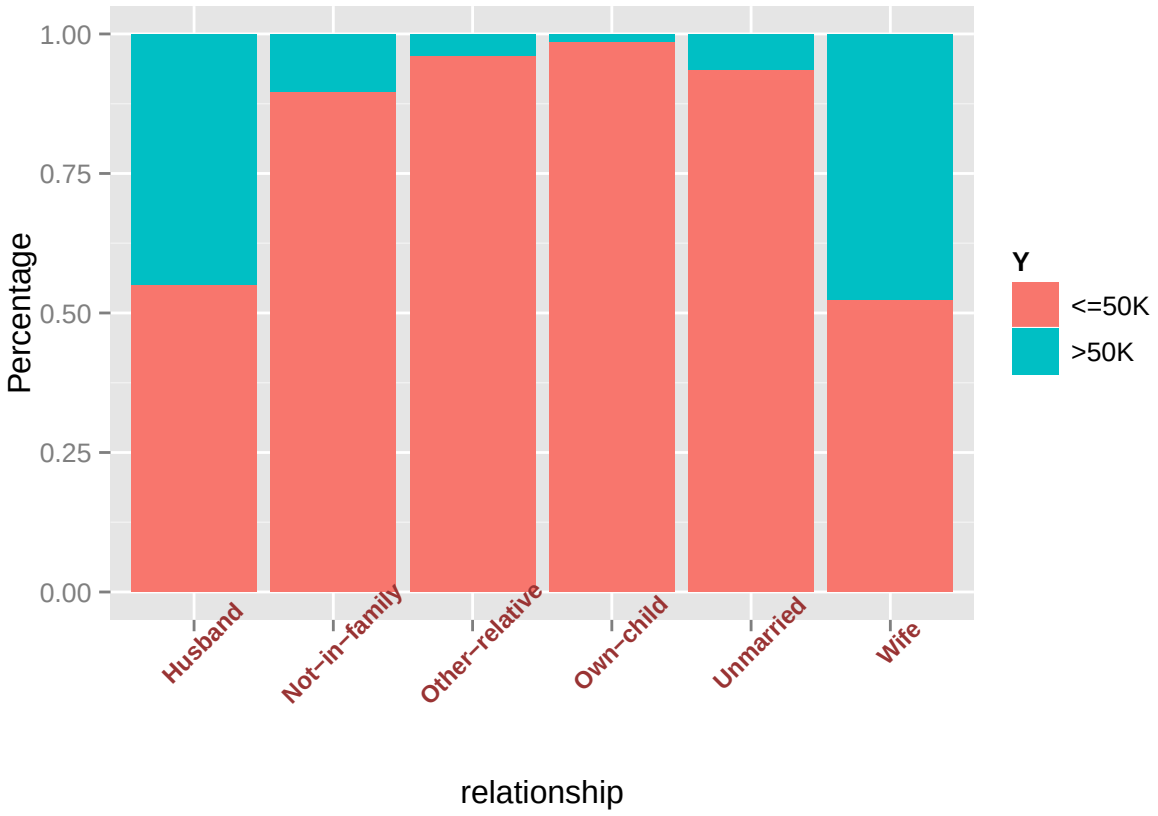
Correlation between two categorcial variables

Use stacked barcharts to find out if outcome Y is affected by [behaves differently across] variables race or relationship. Your code should produce following plots:

```
ggplot(d,aes(x=race,fill=Y))+
  geom_bar(position = "fill")+
  ylab("Percentage")+
  theme(axis.text.x = element_text(face="bold", color="#993333",
                           size=9, angle=45))
```




```
ggplot(d,aes(x=relationship,fill=Y))+
  geom_bar(position="fill")+
  ylab("Percentage")+
  theme(axis.text.x = element_text(face="bold", color="#993333",
                           size=9, angle=45))
```



Note : You'll have to explore option position in geom_bar.

Use smoothing to see natural patterns

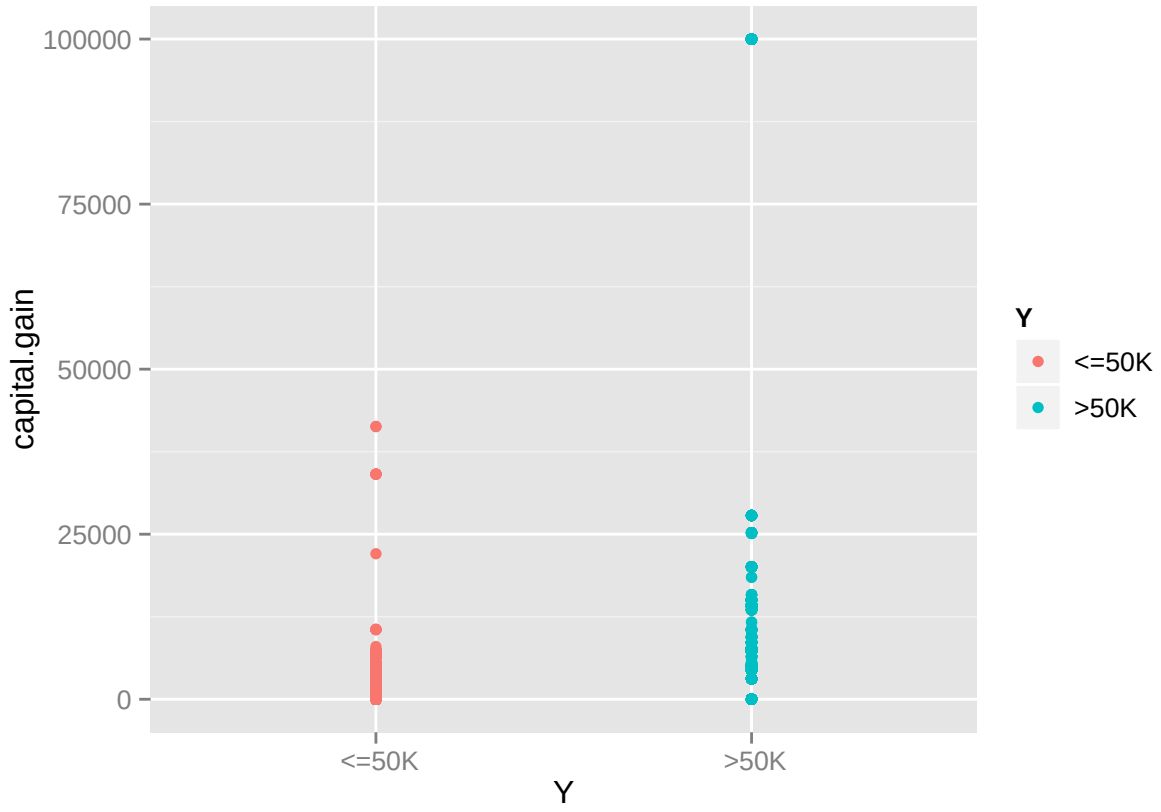
Show visually that hours per week vary similarly across age for both the sexes. Your code should produce following plot:

```
ggplot(d,aes(x=age,y=hours.per.week,color=sex))+
  geom_point(aes(alpha=0.2))+
  geom_smooth()
```

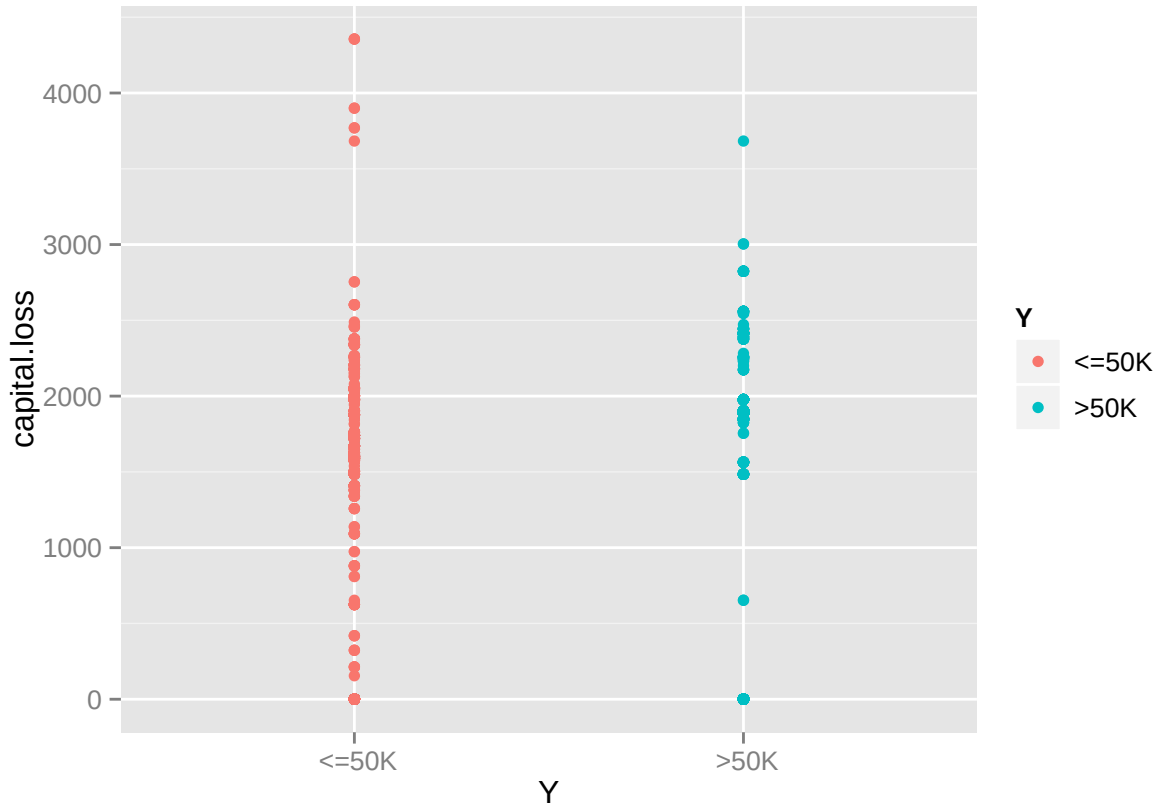
Using point plots with jitter instead of boxplots

Prepare following plots for capital.gain and capital.loss with outcome Y to examine their behaviour across both outcomes.

```
ggplot(d,aes(y=capital.gain,x=Y,color=Y))+
  geom_point()
```

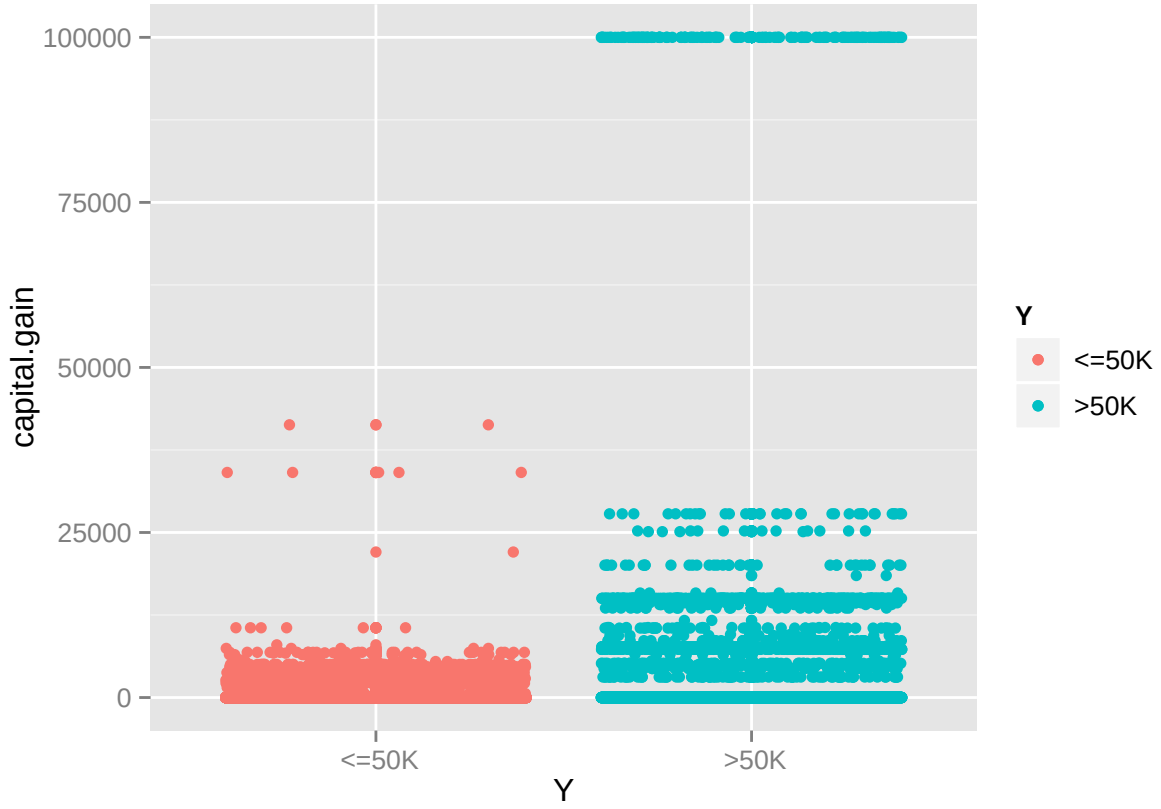



```
ggplot(d,aes(y=capital.loss,x=Y,color=Y))+
  geom_point()
```

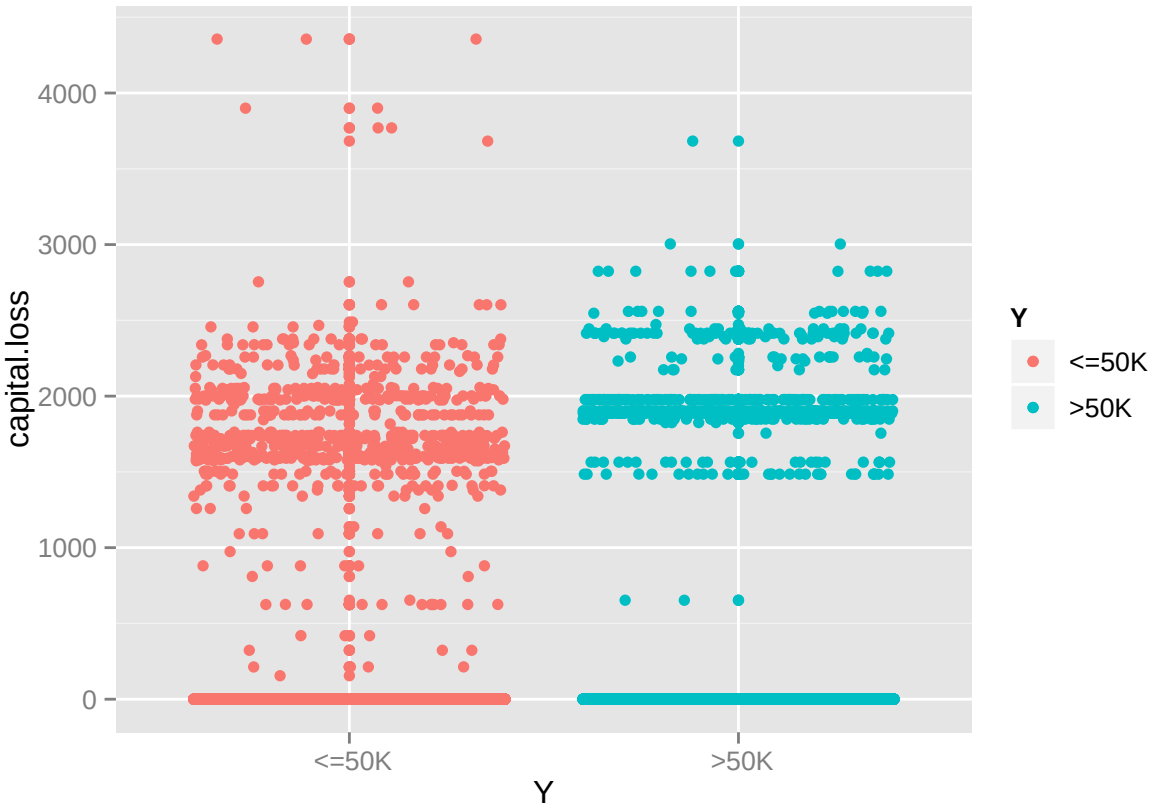


As you can see these plots give a fair idea about ranges of capital.gain and capital.loss but tell nothing about density or count of observations at certain levels which is a crucial set of information. We can remove this drawback by adding some jitter to the point positions. you code should prdocue following plots:

```
ggplot(d,aes(y=capital.gain,x=Y,color=Y))+
  geom_point()+
  geom_jitter()
```



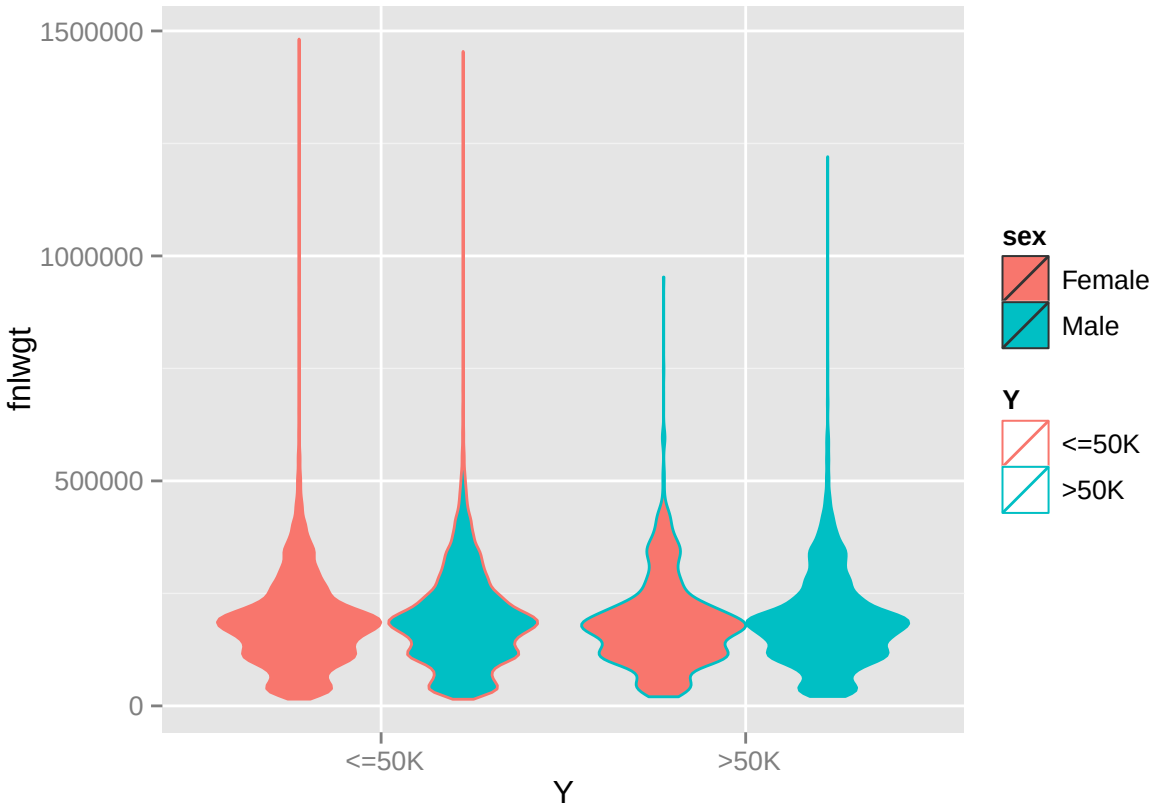

```
ggplot(d,aes(y=capital.loss,x=Y,color=Y))+
  geom_point()+
  geom_jitter()
```



Violin Plots

examine behaviour of fnlwgt for classes of Y [and within that for both the sexes]with . your code should produce following plots:

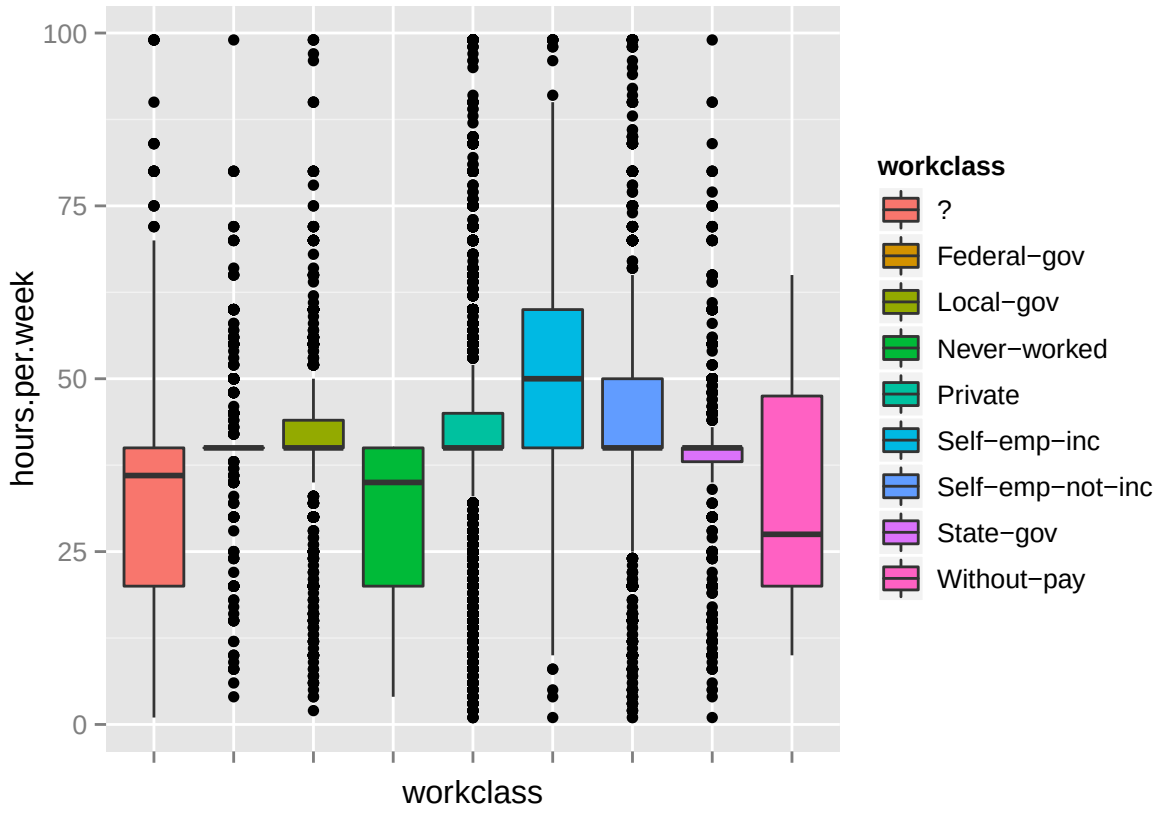
```
ggplot(d,aes(x=Y,y=fnlwgt,color=Y,fill=sex))+
  geom_violin()
```



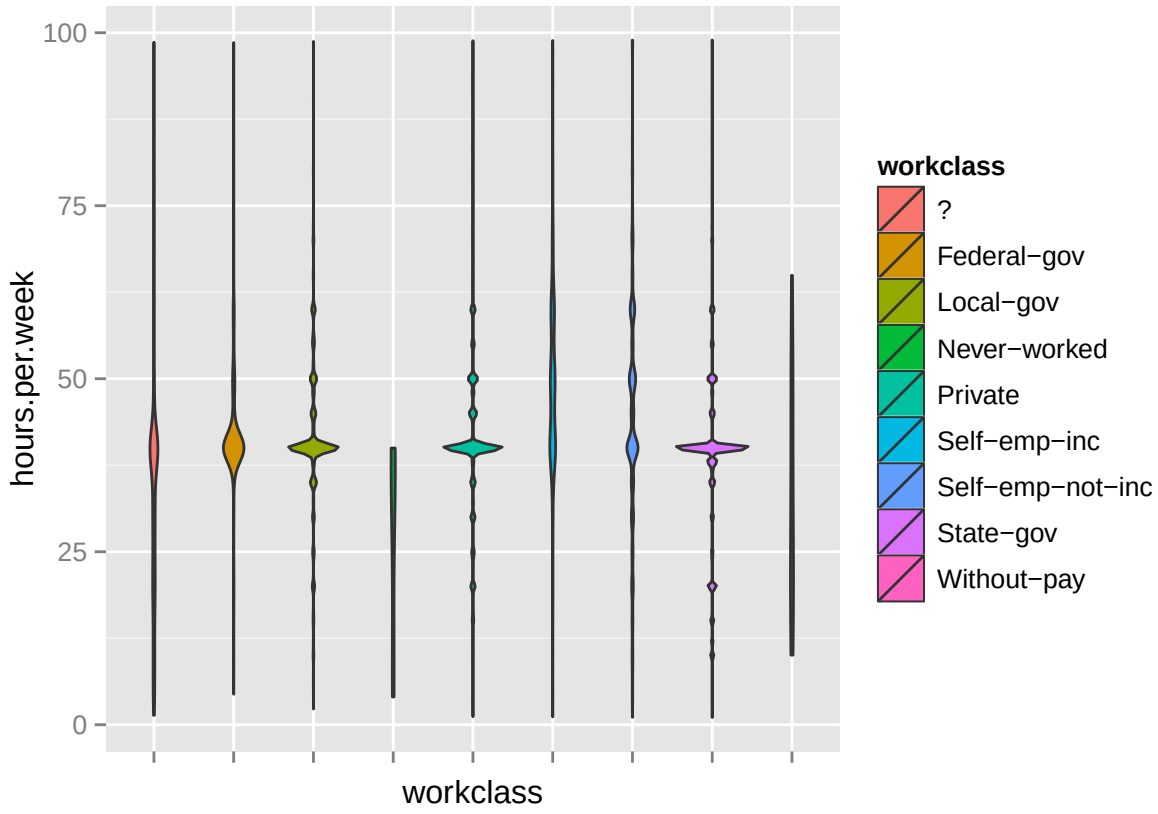
Which visualisation to select

Examine hours.per.week by workclass. which one will be a better choice , a box plot , a violin plot or simple point plot with jitter, think and discuss pros and cons associated with each.

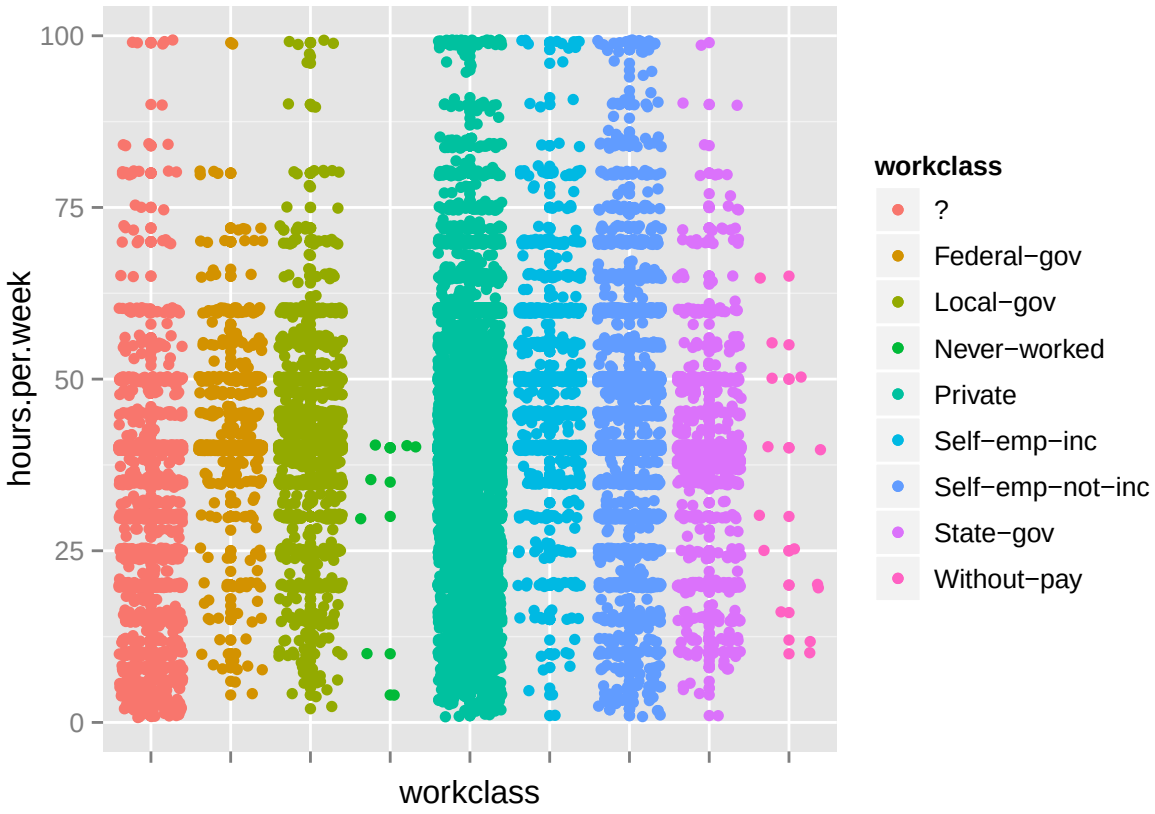
```
ggplot(d,aes(x=workclass,y=hours.per.week,fill=workclass))+
  geom_boxplot()+
  theme(axis.text.x=element_blank())
```




```
ggplot(d,aes(x=workclass,y=hours.per.week,fill=workclass))+
  geom_violin()+
  theme(axis.text.x=element_blank())
```




```
ggplot(d,aes(x=workclass,y=hours.per.week,color=workclass))+
  geom_point()+
  geom_jitter()+
  theme(axis.text.x=element_blank())
```



Discussion on pros and cons of chosing any of these visualisations is left for discussion on the QA forum.