Data set:
Wages data 1 (https://vincentarelbundock.github.io/Rdatasets/doc/plm/Wages.html)

Insight will provide here:

1) **Highlight three descriptive facts from the data with supporting analysis and graphs.**

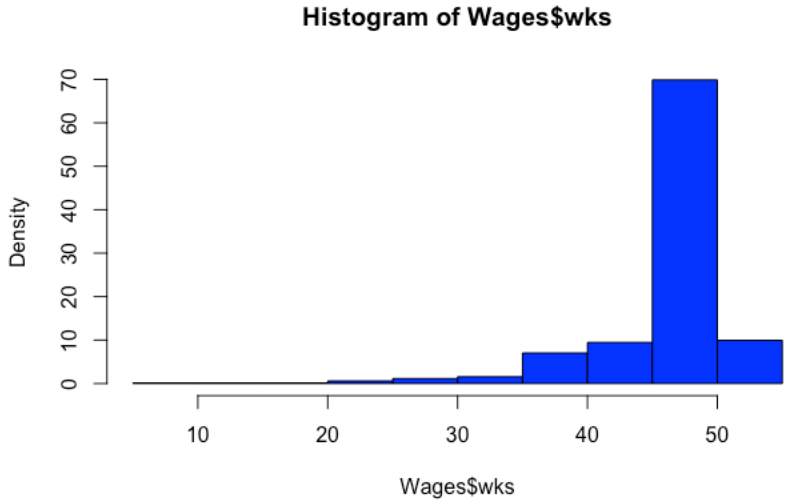
```
library(RCurl)
#When reading data from github, we should pass in the raw version of the data in read.csv(),
#We should get the URL for the raw version by clicking on the Raw button displayed above the data.
x <- getURL("https://raw.githubusercontent.com/jazzsun000/how-to-build-statistical-distribution-analysis-on-wage-data/master/Wages.csv")
Wages <- read.csv(text = x)

summary(Wages)
```

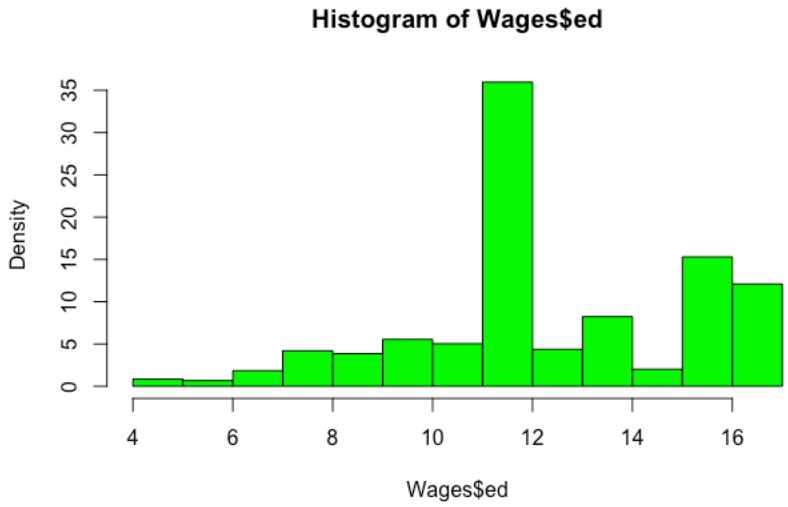
```
       X              exp              wks        bluecol         ind          south
 Min.   :   1   Min.   : 1.00   Min.   : 5.00   no :2036   Min.   :0.0000   no :2956
 1st Qu.:1042   1st Qu.:11.00   1st Qu.:46.00   yes:2129   1st Qu.:0.0000   yes:1209
 Median :2083   Median :18.00   Median :48.00              Median :0.0000
 Mean   :2083   Mean   :19.85   Mean   :46.81              Mean   :0.3954
 3rd Qu.:3124   3rd Qu.:29.00   3rd Qu.:50.00              3rd Qu.:1.0000
 Max.   :4165   Max.   :51.00   Max.   :52.00              Max.   :1.0000
  smsa       married        sex        union          ed        black          lwage
 no :1442   no : 773   female: 469   no :2649   Min.   : 4.00   no :3864   Min.   :4.605
 yes:2723   yes:3392   male  :3696   yes:1516   1st Qu.:12.00   yes: 301   1st Qu.:6.395
                                                Median :12.00              Median :6.685
                                                Mean   :12.85              Mean   :6.676
                                                3rd Qu.:16.00              3rd Qu.:6.953
                                                Max.   :17.00              Max.   :8.537
> mean(Wages$exp)
```

1.

From desity histogram of years of full-time work experience data, we can find there are some descriptive facts here.
Overall, over **30%** of the working experience fall under **5 to 15 years**.



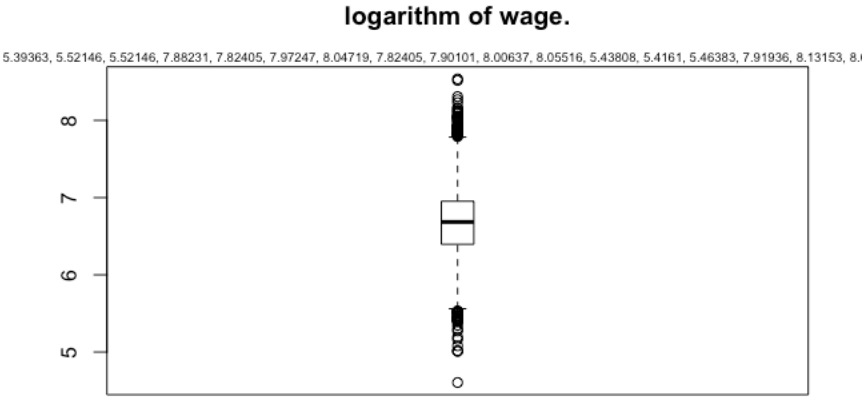
As for weeks worked, around **70%** of people weeks they worked fall under **45 to 50 weeks.**



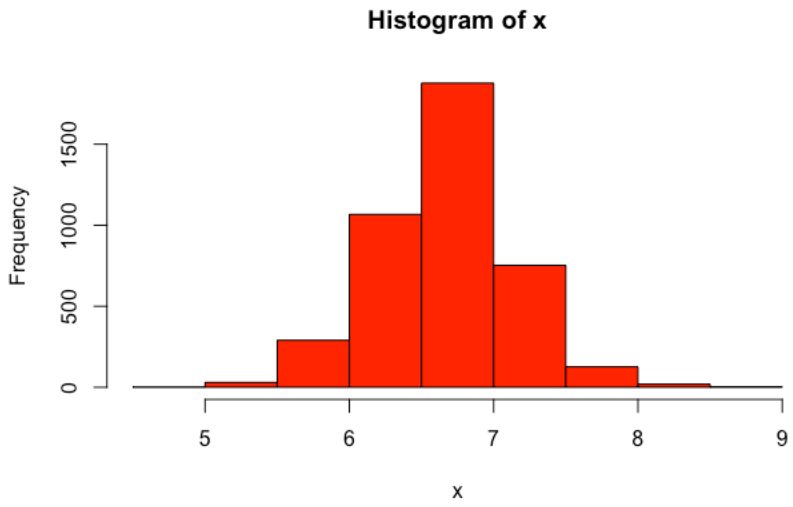
If we take a look on their years of education, 11 to 12 years is the most. It account for 35%.

2) **Pick a continuous variable of interest, what is the distribution of this variable?**
3) **Continuing from above, how would you examine and clean outlier?**

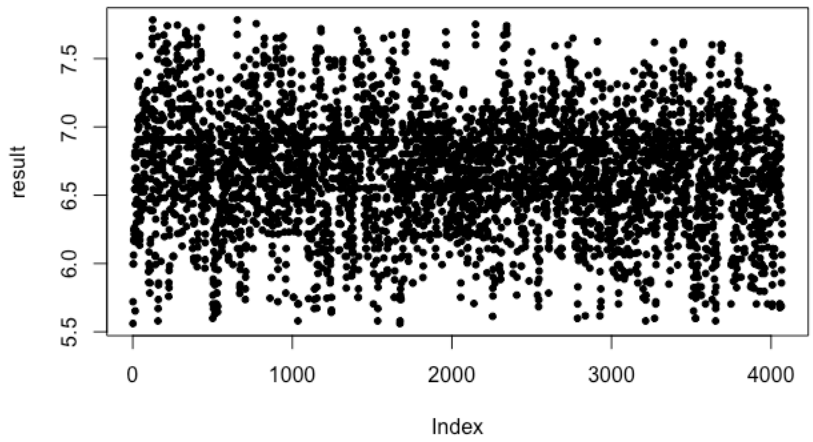
#clean outlier
#For a given continuous variable, outliers are those observations that lie outside 1.5 * IQR, where IQR, the 'Inter Quartile Range' is the difference between 75th and 25th quartiles. Look at the points outside the whiskers in below box plot
outlier_values <- boxplot.stats(Wages$lwage)$out # outlier values.
boxplot(Wages$lwage, main="logarithm of wage.", boxwex=0.1)
mtext(paste("Outliers: ", paste(outlier_values, collapse=", ")), cex=0.6)
here are outliers



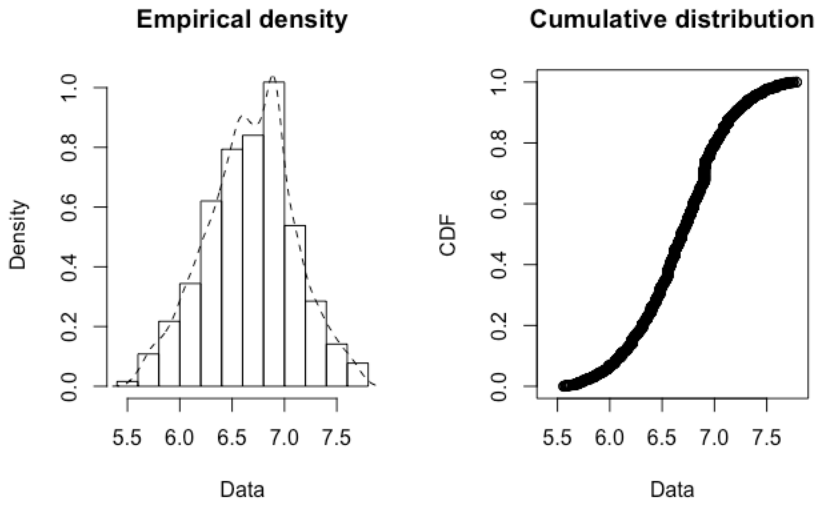
```
x<-Wages$lwage
result = x[!x %in% boxplot.stats(x)$out]
hist(x, breaks=12, col="red")
hist(result, breaks=12, col="blue")
```



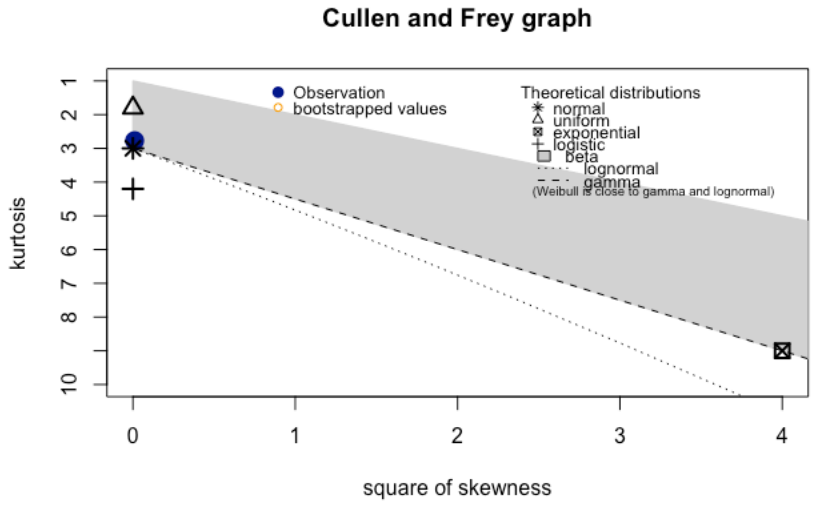

```
library(fitdistrplus)
plot(result, pch=20)
```



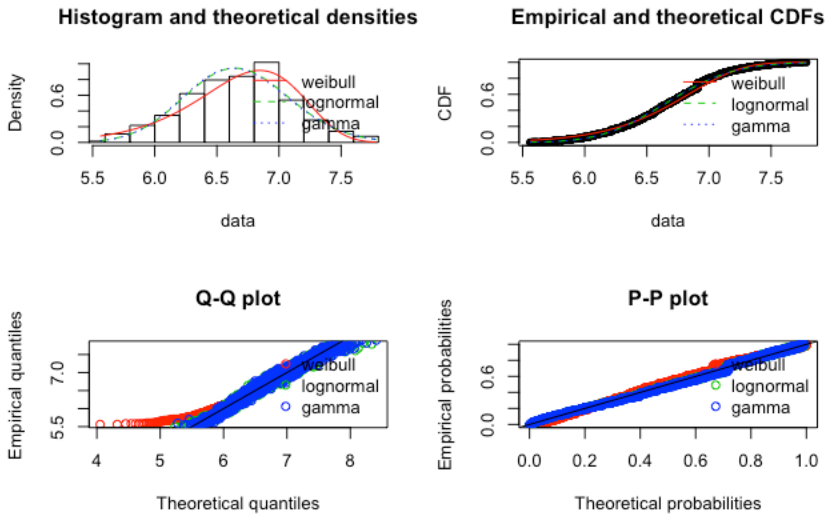

```
plotdist(result, histo = TRUE, demp = TRUE)
```



descdist(result, discrete=FALSE, boot=500)




```
#Say, in the previous eg, we chose the weibull, gamma and log-normal to fit:
fit_w  <- fitdist(result, "weibull")
fit_g  <- fitdist(result, "gamma")
fit_ln <- fitdist(result, "lnorm")
summary(fit_ln)
#we can plot the results:
par(mfrow=c(2,2))
plot.legend <- c("weibull", "lognormal", "gamma")
denscomp(list(fit_w, fit_g, fit_ln), legendtext = plot.legend)
cdfcomp (list(fit_w, fit_g, fit_ln), legendtext = plot.legend)
qqcomp  (list(fit_w, fit_g, fit_ln), legendtext = plot.legend)
ppcomp  (list(fit_w, fit_g, fit_ln), legendtext = plot.legend)
```

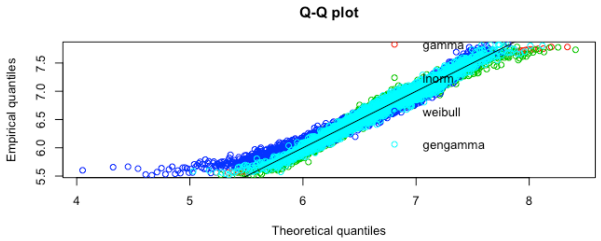


As a result ,we need to test which distribution it belong to? weibull?lognormal?or gamma?
From plot above, it look like more close to gamma distribution

```
install.packages(flexsurv)
library(flexsurv) # on CRAN

gengammafit  <-  fitdistrplus::fitdist(result, "gengamma",
                                       start=function(d) list(mu=mean(d),
                                                              sigma=sd(d),
                                                              Q=0))

qqcomp(list(fit_g, fit_ln, fit_w, gengammafit),legendtext=c("gamma", "lnorm", "weibull", "gengamma") )
```



But if take a further more look on above plot.

None of the distributions fit very well in the right (upper) tail, but the generalized gamma is best

```
> gengammafit$aic
[1] 4511.414
> fit_w$aic
[1] 4795.813
> fit_g$aic
[1] 4544.802
> fit_ln$aic
[1] 4565.646
```

#A good model is the one that has minimum AIC among all the other models.
#And if we take a look from AIC, generalized gamma distribution show better than the other model.