Datas - cars characteristics.

First column- car model. Second - amount miles per gallon. Third - amount of cylinders. Four - engime volume. Five - Power. Six - weight car. Seven - production year.

№	Cars	Miles Per gallon	Cylinders	Engine volume	Power	Weight	Year
1	Mazda RX4	21.0	6	160.0	110	2.620	1999
2	Mazda RX4 Wag	21.0	6	160.0	110	2.875	2008
3	Datsun 710	22.8	4	108.0	93	2.320	2008
4	Hornet 4 Drive	21.4	6	258.0	110	3.215	1999
5	Hornet Sportabout	18.7	8	360.0	175	3.440	2008
6	Valiant	18.1	6	225.0	105	3.460	2008
7	Duster 360	14.3	8	360.0	245	3.570	2008
8	Merc 240D	24.4	4	146.7	62	3.190	2008
9	Merc 230	22.8	4	140.8	95	3.150	1999
10	Merc 280	19.2	6	167.6	123	3.440	2008
11	Merc 280C	17.8	6	167.6	123	3.440	1999
12	Merc 450SE	16.4	8	275.8	180	4.070	2008
13	Merc 450SL	17.3	8	275.8	180	3.730	1999
14	Merc 450SLC	15.2	8	275.8	180	3.780	1999
15	Cadillac Fleetwood	10.4	8	472.0	205	5.250	1999
16	Lincoln Continental	10.4	8	460.0	215	5.424	1999
17	Chrysler Imperial	14.7	8	440.0	230	5.345	2008
18	Fiat 128	32.4	4	78.7	66	2.200	1999
19	Honda Civic	30.4	4	75.7	52	1.615	1999
20	Toyota Corolla	33.9	4	71.1	65	1.835	2008
21	Toyota Corona	21.5	4	120.0	97	2.465	1999
22	Dodge Challenger	15.5	8	318.0	150	3.520	2008
23	AMC Javelin	15.2	8	304.0	150	3.435	2008
24	Camaro Z28	13.3	8	350.0	245	3.840	2008
25	Pontiac Firebird	19.2	8	400.0	175	3.845	2008
26	Fiat X1-9	27.3	4	79.0	66	1.935	1999
27	Porsche 914-2	26.0	4	120.3	91	2.140	2008
28	Lotus Europa	30.4	4	95.1	113	1.513	1999
29	Ford Pantera L	15.8	8	351.0	264	3.170	1999
30	Ferrari Dino	19.7	6	145.0	175	2.770	2008
31	Maserati Bora	15.0	8	301.0	335	3.570	2008
32	Volvo 142E	21.4	4	121.0	109	2.780	1999

characteristic of data:

```
str(data)
```

```
## 'data.frame'       : 32 obs. of  7 variables:
##  $ Cars             : Factor w/ 32 levels "AMC Javelin",..: 18 19 5 13 14 31...
##  $ Miles.per.gallon: num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ Cylinders        : int  6 6 4 6 8 6 8 4 4 6 ...
##  $ Volume           : num  160 160 108 258 360 ...
##  $ Power            : int  110 110 93 110 175 105 245 62 95 123 ...
##  $ Weight           : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ Year             : int  1999 2008 2008 1999 2008 2008 2008 2008 1999 2008 ...
```

Amount of cars that have 4,6,7 cylinders:

```
table(data$Cylinders)
```

```
##  4  6  8
## 11  7 14
```

```
Amount of cars that were produced in 1999 and 2008 years.
```

```
 table(data$Year)
```

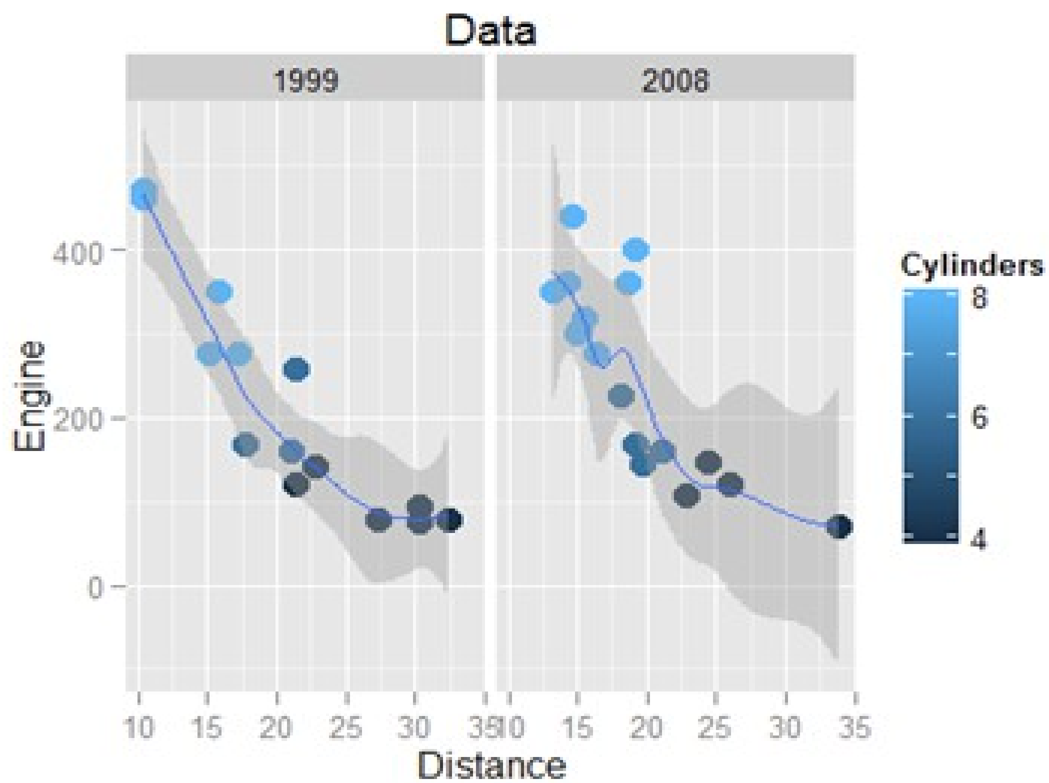
```
## 1999 2008
##   15   17
```

Or in general:

```
table(data$Cylinders,data$Year)
```

```
##     1999 2008
##   4    7    4
##   6    3    4
##   8    5    9
```

Lets build the plots:



More cylinders are, more powerful machine. And consumes more fuel.

Tests on normality:

1. The Anderson-Darling Test.
2. The Shapiro-Wilk Test.

Null hypothesis is that data is close to normal distribution, alternative - not

Reject null hypothesis when p<0.05. When p>0.05 accept.

Check on normality first column mpg (Miles per gallon)

```
shapiro.test(data$Miles.per.gallon)
```

```
##  Shapiro-Wilk normality test
##
## data:  data$Miles.per.gallon
## W = 0.9476, p-value = 0.1229
```

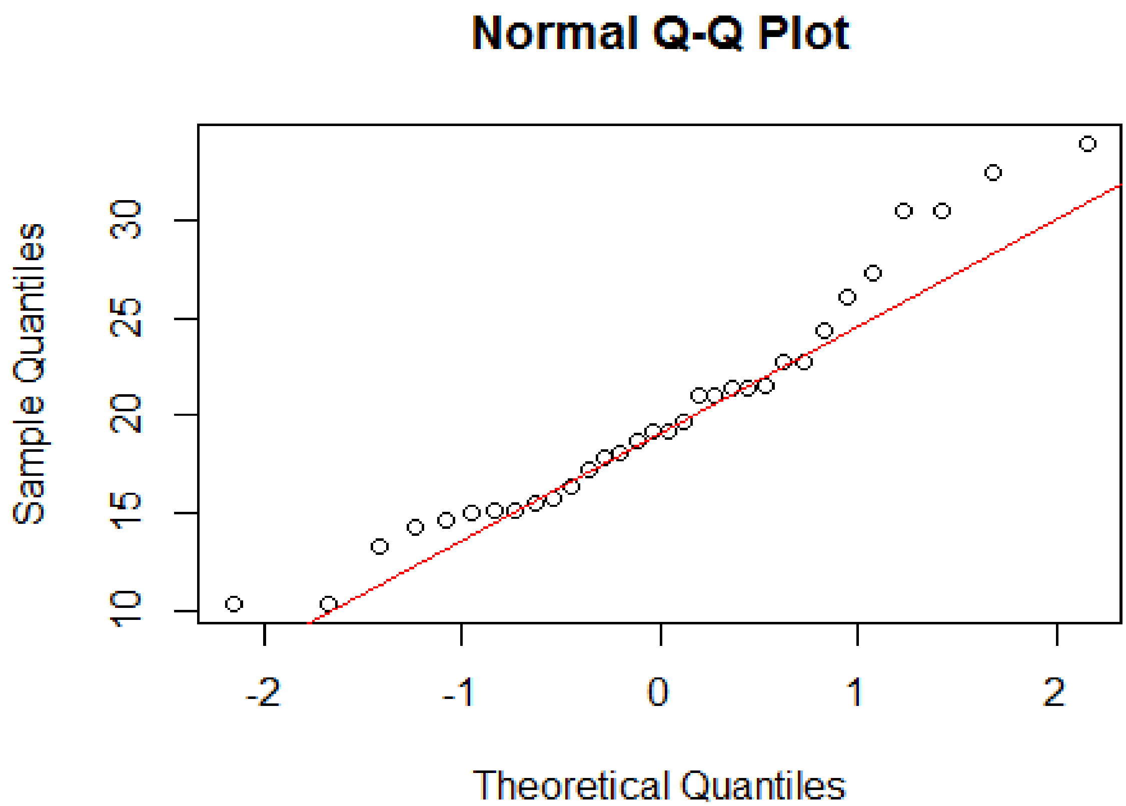
```
ad.test(data$Miles.per.gallon)
```

```
##  Anderson-Darling normality test
##
```

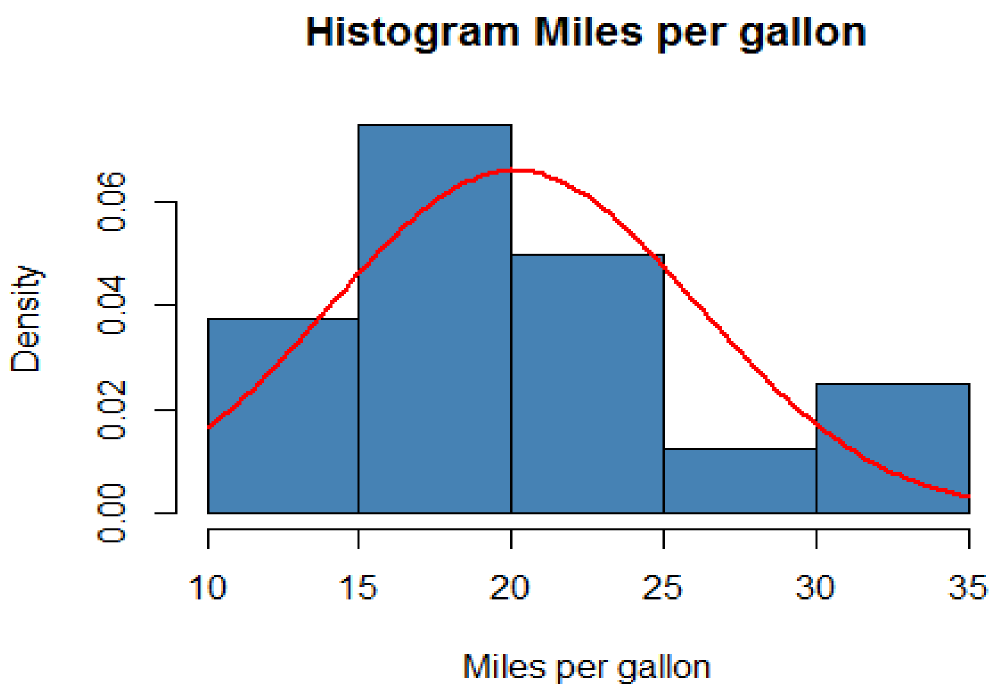
```
## data:  data$Miles.per.gallon
## A = 0.5797, p-value = 0.1207
```

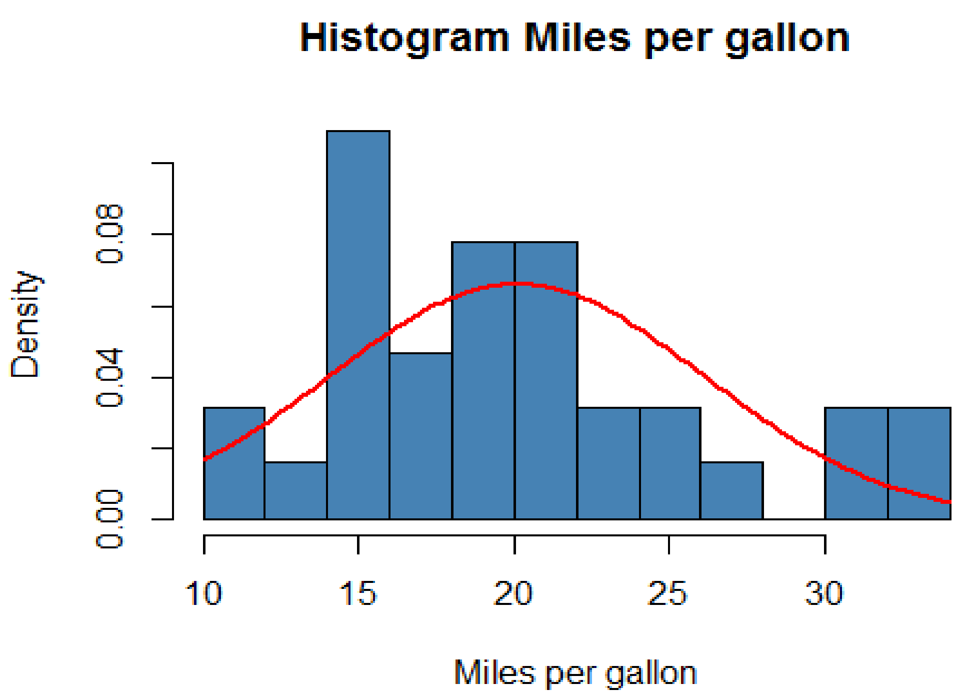
We can see that p<0.05, that's why reject null hypothesis.

Let's see it on the plots.



Easy to see, that distribution is not normal.





Let's perform similar calculations for the next column data

“Engine Volume”.

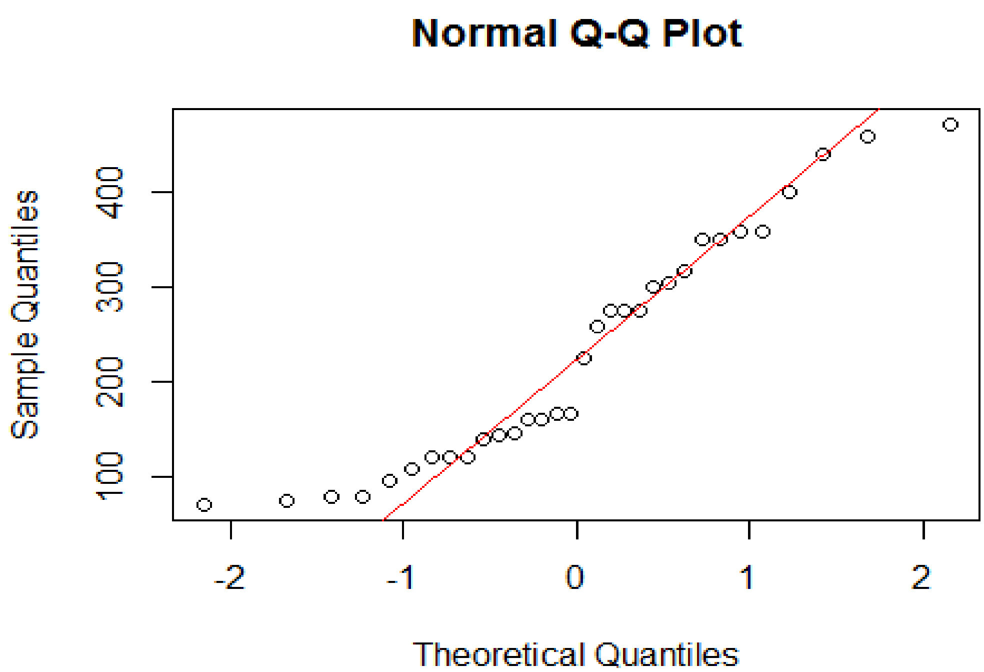
```
shapiro.test(data$Volume)
```

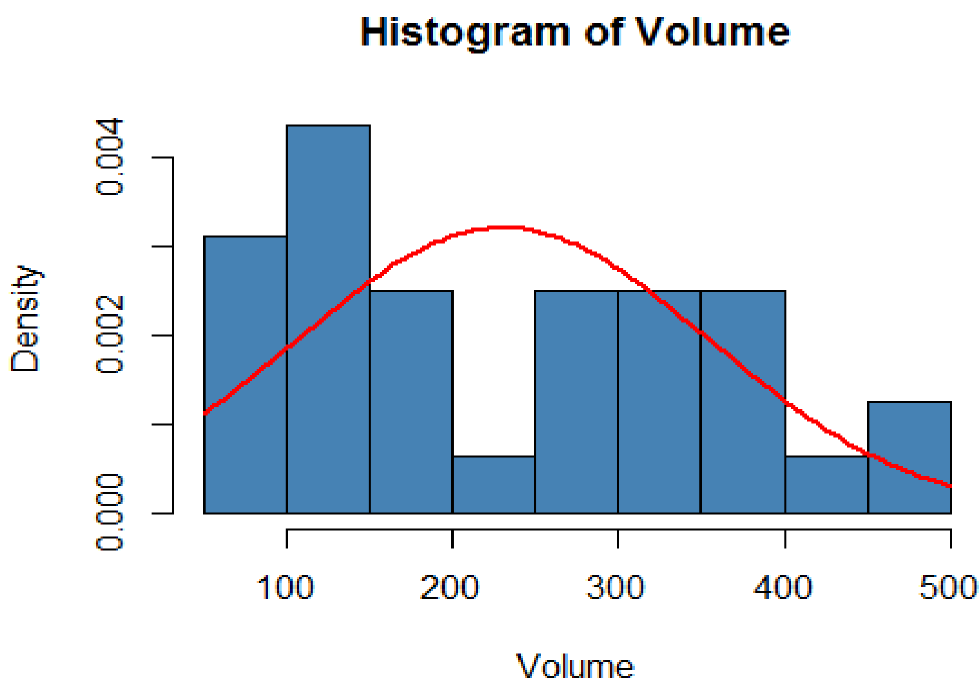
```
##  Shapiro-Wilk normality test
##
## data:  data$Volume
## W = 0.92, p-value = 0.02081
```

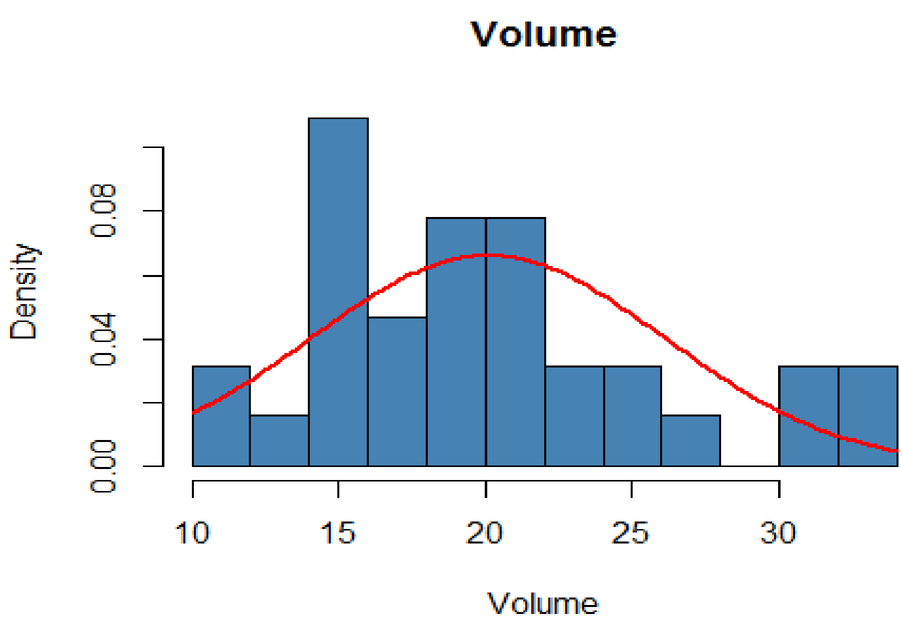
```
ad.test(data$Volume)
```

```
##
##  Anderson-Darling normality test
##
## data:  data$Volume
## A = 0.8745, p-value = 0.02211
```

We can see that $p<0.05$, that’s why reject null hypothesis.





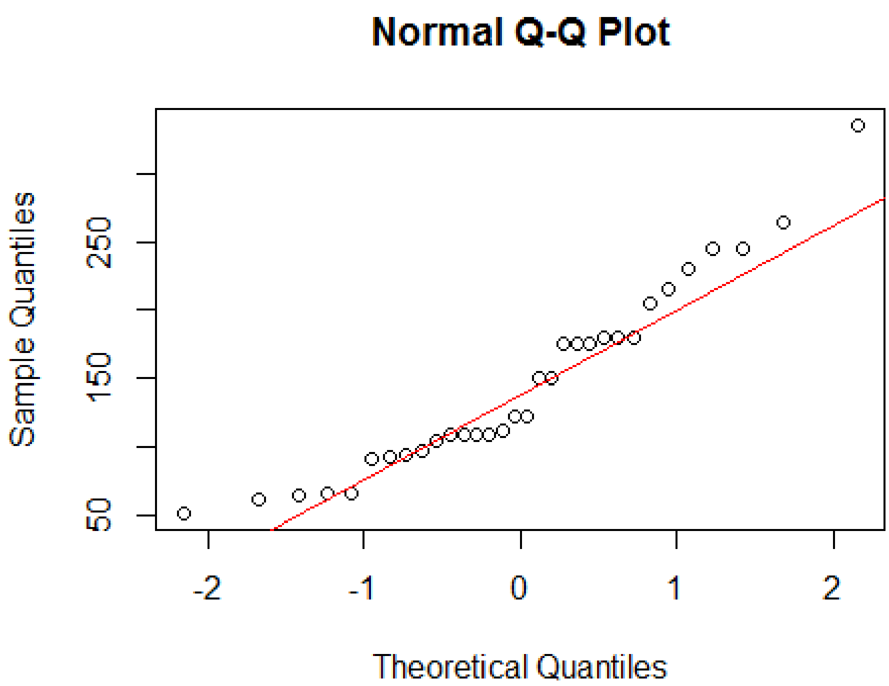


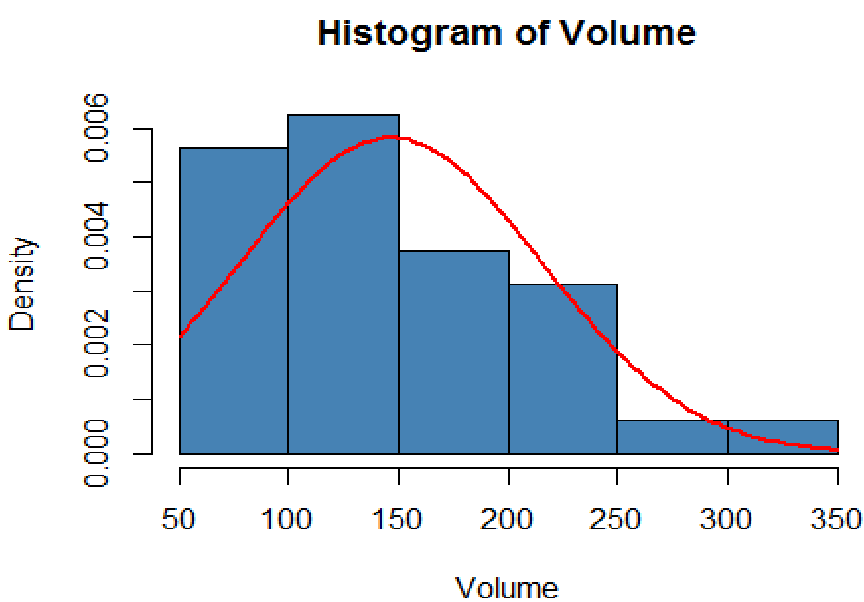
Let's perform similar calculations for the next column data:

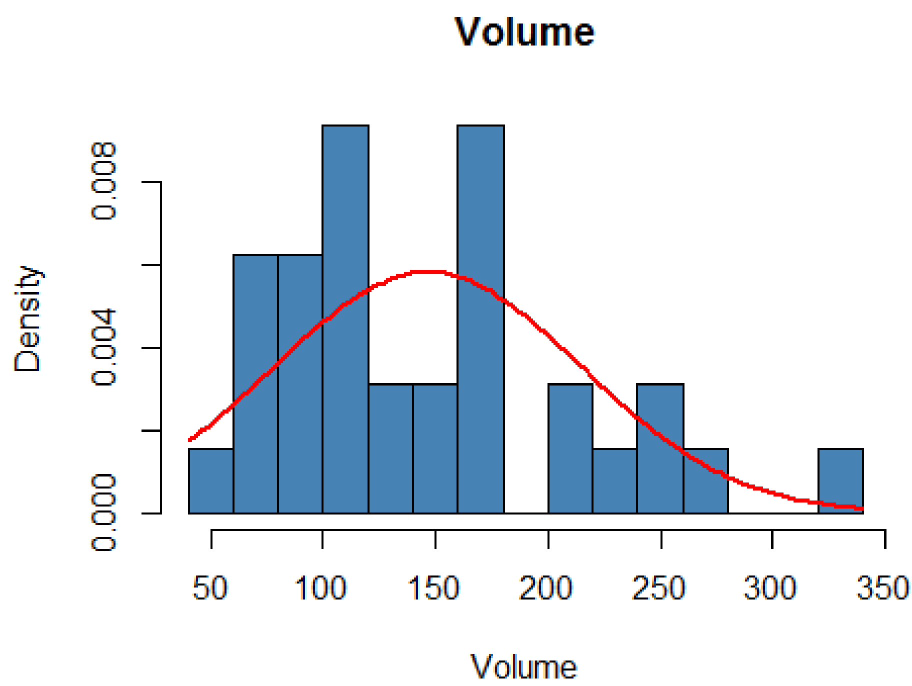
```r
shapiro.test(data$Power)
```

```
##  Shapiro-Wilk normality test
##
## data:  data$Power
## W = 0.9334, p-value = 0.04881
```

We can see that $p<0.05$, that’s why reject null hypothesis.


Normal Q-Q Plot
Sample Quantiles
50
150
250
-2
-1
0
1
2
Theoretical Quantiles


Histogram of Volume
Density
0.000
0.002
0.004
0.006
50
100
150
200
250
300
350
Volume

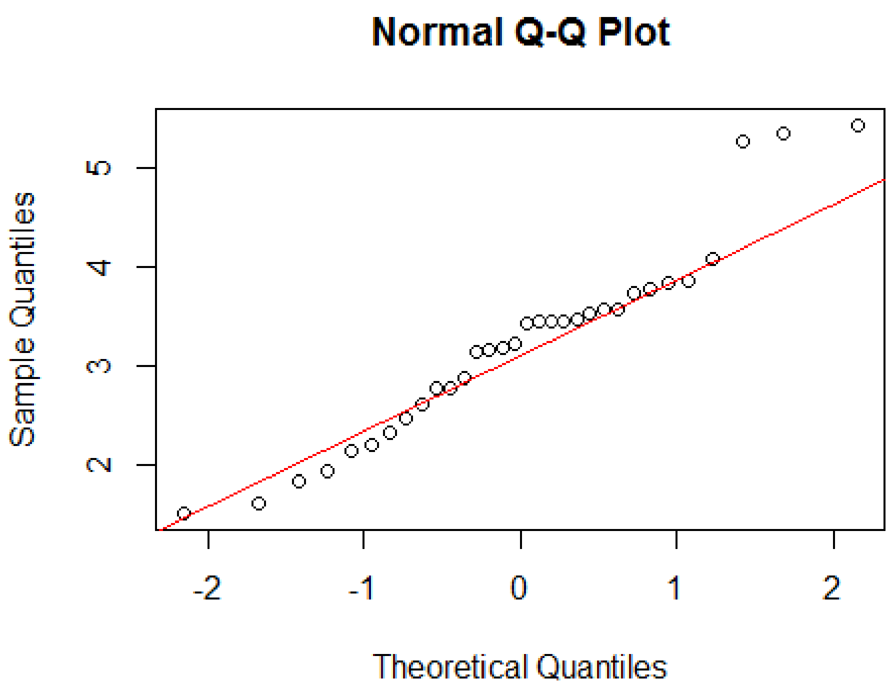


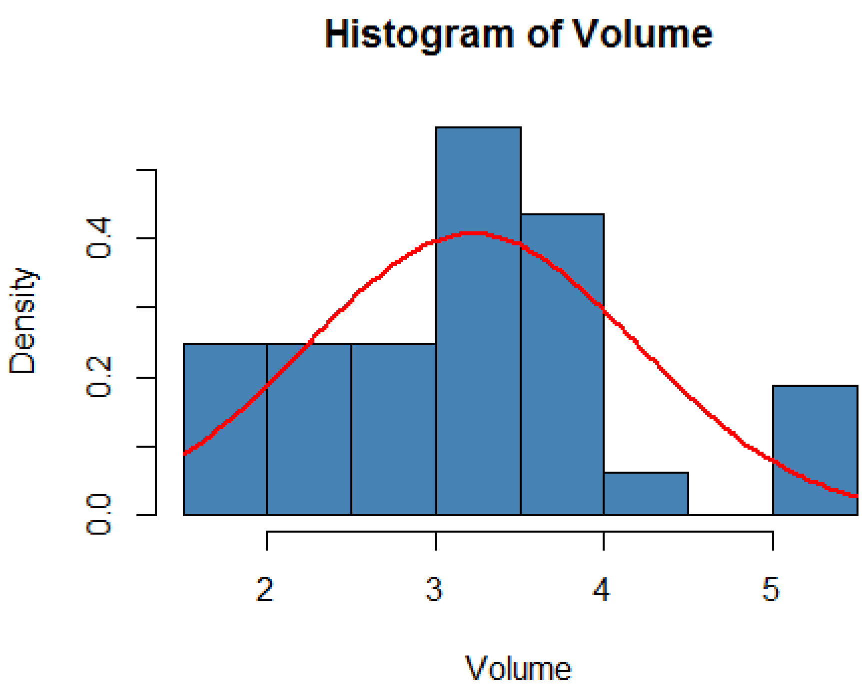
Let's analyze data from "Weight" column.

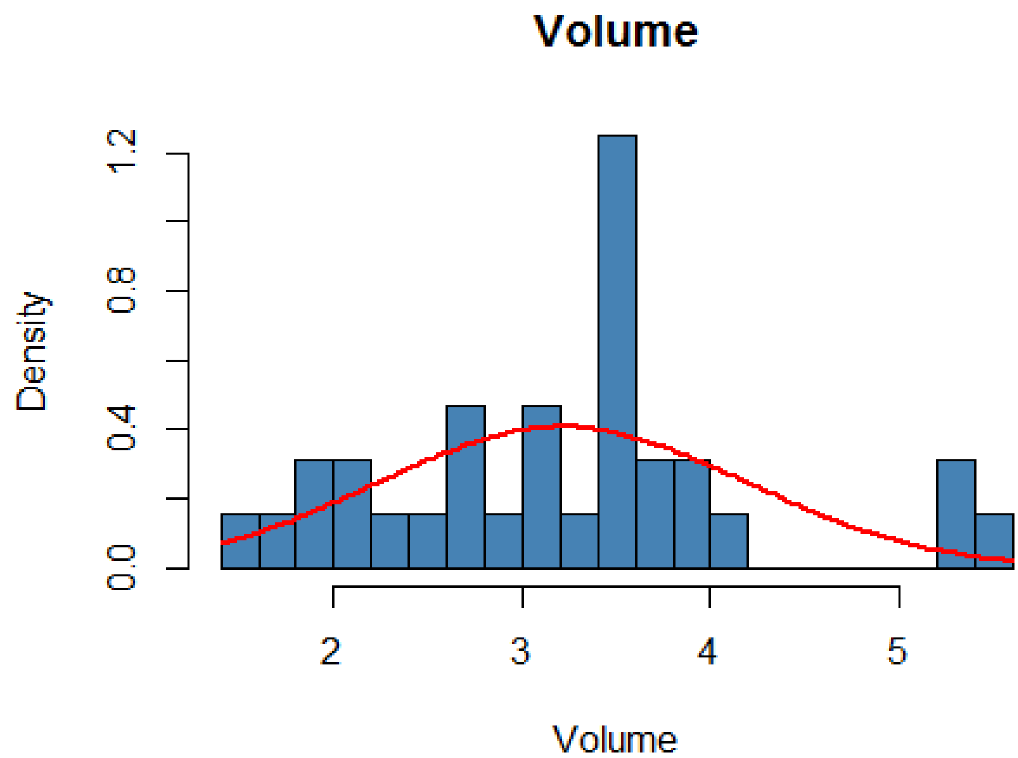
```
shapiro.test(data$Weight)

##
##  Shapiro-Wilk normality test
##
## data:  data$Weight
## W = 0.9433, p-value = 0.09265
```

```
ad.test(data$Weight)

##
##  Anderson-Darling normality test
##
## data:  data$Weight
## A = 0.6091, p-value = 0.1038
```

We can see that p>0.05, that's why accept null hypothesis.


Normal Q-Q Plot
Sample Quantiles
2
3
4
5
-2
-1
0
1
2
Theoretical Quantiles


Histogram of Volume
Density
0.0
0.2
0.4
2
3
4
5
Volume



Let’s use t-test for comparing mpg means, and cars with 4 and 6 cylinders.

Null hypotheses that two means are equal.

```
d1<-data$Miles.per.gallon[data$Cylinders==4]    ## cars with 4 cylinders

d2<-data$Miles.per.gallon[data$Cylinders==6]    ## cars with 6 cylinders
t.test(d1,d2)


##  Welch Two Sample t-test
##
## data:  d1 and d2
## t = 4.7191, df = 12.956, p-value = 0.0004048
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.751376 10.090182
## sample estimates:
## mean of x mean of y
##  26.66364  19.74286
```

So, we can see p is very small (**p-value = 0.0004048).** That’s why reject Null hypothesis.

Let’s use t-test for comparing “Engine Volume” with cars that were produced in 1999, 2008 years.

$H0: \mu1 = \mu2$ або $\mu1 - \mu2 = 0$

$H_1: \mu_1 \neq \mu_2$ або $\mu_1 - \mu_2 \neq 0$

```
d1<-data$Volume[data$Year==1999]  ## 1999 year
d2<-data$Volume[data$Year==2008]  ## 2008 year
t.test(d1,d2)
```

```
##  Welch Two Sample t-test
##
## data:  d1 and d2
## t = -0.9326, df = 27.731, p-value = 0.3591
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -132.5053   49.6245
## sample estimates:
## mean of x mean of y
##  208.7067  250.1471
```

We can see p> 0.05 (**p-value = 0.3591)**, accept null hypothesis.

We can make a conclusion, that only data from "Weight" column have distribution close to normal. Plots show us that data is natural and display real results. Also, t-test of cars volume showed us positive result.