Assignment 1 | Exercise 1

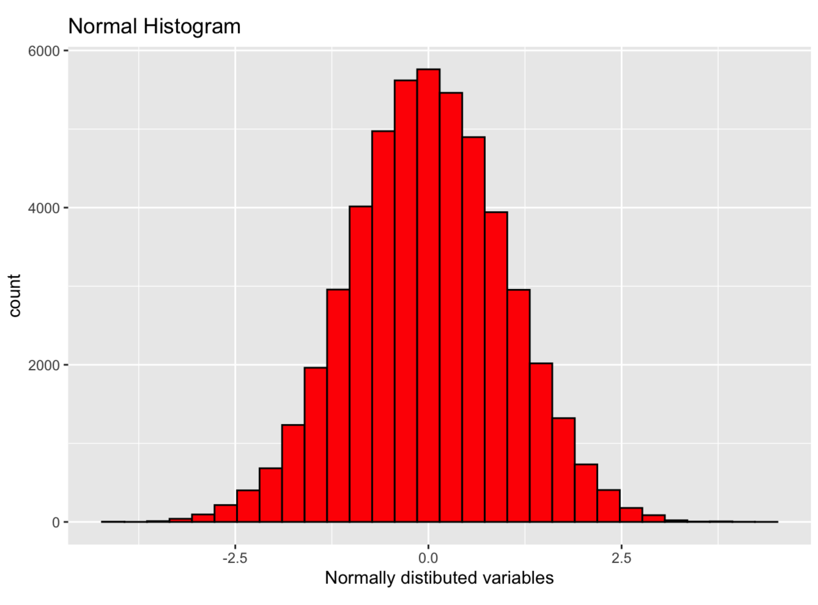
2022-06-16

1.

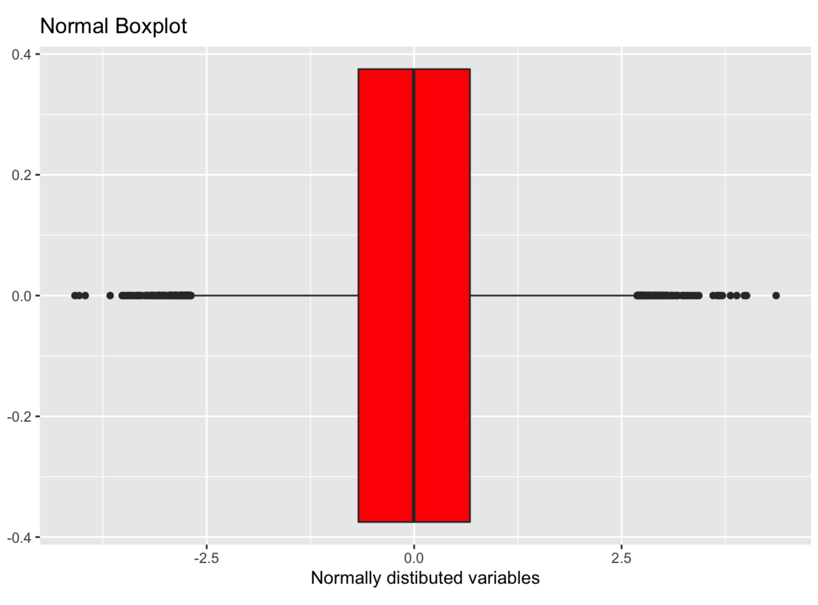
```
library(Pareto)
library(ggplot2)
library(gridExtra)

set.seed(100)
Data = data.frame(x.n=rnorm(50000),x.p=rPareto(50000,t=1,alpha=2))


summary(Data$x.n)
```

```
##       Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -4.087893 -0.671144 -0.005919 -0.000208  0.672466  4.363243
```

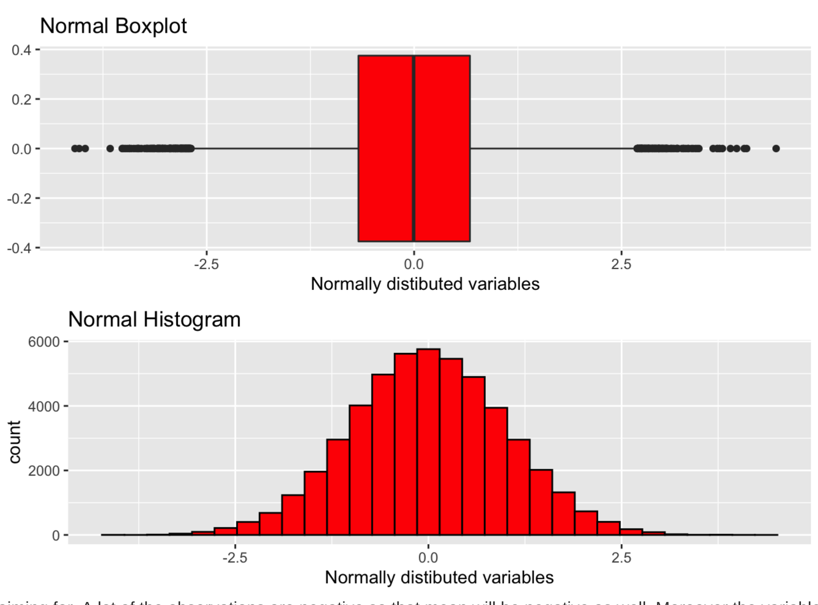
```
H=ggplot(data = Data) + aes(x= x.n) + geom_histogram(color = "black", fill = "red",bins = 30) + ggtitle("Normal H
istogram") + xlab("Normally distibuted variables")
H
```




```
B=ggplot(data = Data) + aes(x = x.n) + xlab("Normally distibuted variables") + geom_boxplot(fill = "red") + ggtit
le("Normal Boxplot")
B
```




```
grid.arrange(B,H, nrow = 2)
```



2. Yes, it is basically what we were aiming for. A lot of the observations are negative so that mean will be negative as well. Moreover the variables are standard normally distributed from the fact that mean is around 0 and standard deviation is around zero.

```
#Mean=Median=Mode
nmean =mean(Data$x.n) #-0.0002084956
nmean
```

```
## [1] -0.0002084956
```

```
nsd = sd(Data$x.n) #0.9989658
nsd
```

```
## [1] 0.9989658
```

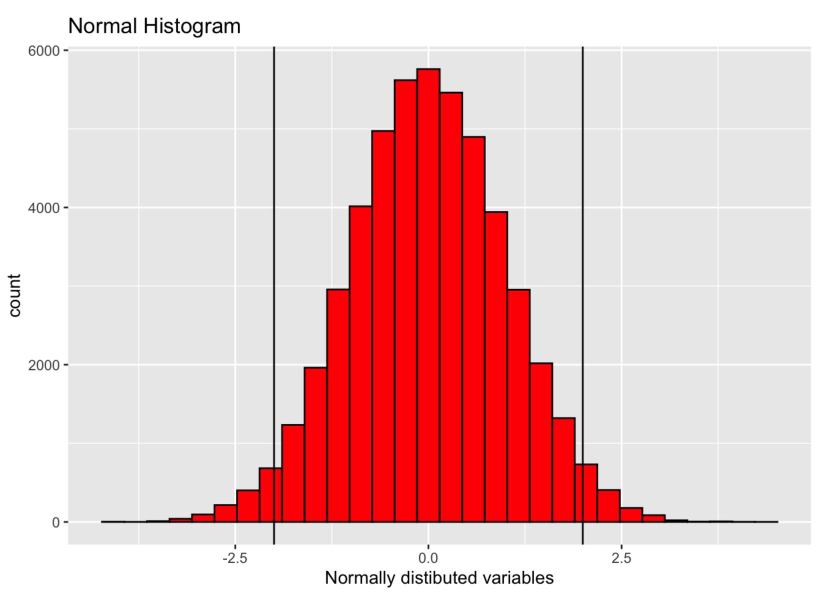
3. Once summarized the interval it can be seen that 47664 observations lay inside the interval mentioned in the code. Following that summarizing the data is possible and will be exact.

Looking on the graph and the console outcome, we can conclude that most of the observations are included in the interval mentioned (MEAN -2*SD).Following that summarizing the data is possible and the mean can be used to predict new observations. Moreover variable is standard normally distributed what can be seen after checking mean and standard deviation.

```
summarize_interval<- subset(Data$x.n, Data$x.n > nmean - 2* nsd & Data$x.n< nmean + 2* nsd)
length(summarize_interval)
```

```
## [1] 47664
```

```
S=ggplot(data = Data) + aes(x= x.n) + geom_histogram(color = "black", fill = "red",bins = 30) +ggtitle("Normal Hi
stogram") + xlab("Normally distibuted variables") + geom_vline(xintercept= nmean - 2* nsd) + geom_vline(xintercep
t = nmean + 2*nsd)
S
```



4.

pmean = 1.993904 psd = 2.601173

Once found mean and standard deviation, we moved to looking for the outliers. That procedure helped us see how many observations might be the extreme ones. 45302 is the number of observations laying inside the interval. Especially once we plot both distributions we can see that there are outliers appearing. The mean neglects new very extreme realizations because extreme realizations will lay outside the interval shown on the histogram. Moreover the mean and standard deviation is bad predictor from the fact that data on the histogram is not placed symmetrically and most of the observations are laying below the mean line. It can be easily assumed that mean is not the best predictor. “plot(Data$x.n, Data$x.p)” also by plotting both sets of variables outliers can be seen.

I did not use filter function since removed outliers in the way that there was no need to filter them out.

```
pmean = mean(Data$x.p) #1.993904

psd = sd(Data$x.p) #2.601173


Q1 <- quantile(Data$x.p,.025)
Q1
```

```
##     2.5%
## 1.012447
```

```
Q3 <- quantile(Data$x.p,.75)
Q3
```

```
##      75%
## 1.991712
```

```
IQR <- IQR(Data$x.p)
IQR
```

```
## [1] 0.8373718
```

```
no_outliers <- subset(Data$x.p, Data$x.p > (Q1 - 1.5*IQR) & Data$x.p < (Q3 + 1.5*IQR))
length(no_outliers)
```

```
## [1] 45302
```

```
summary(no_outliers)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.138   1.351   1.528   1.763   3.248
```

```
summarize_pareto<- subset(Data$x.p, Data$x.p > pmean - 2* psd & Data$x.p< pmean + 2* psd)
length(summarize_pareto)    #49036
```

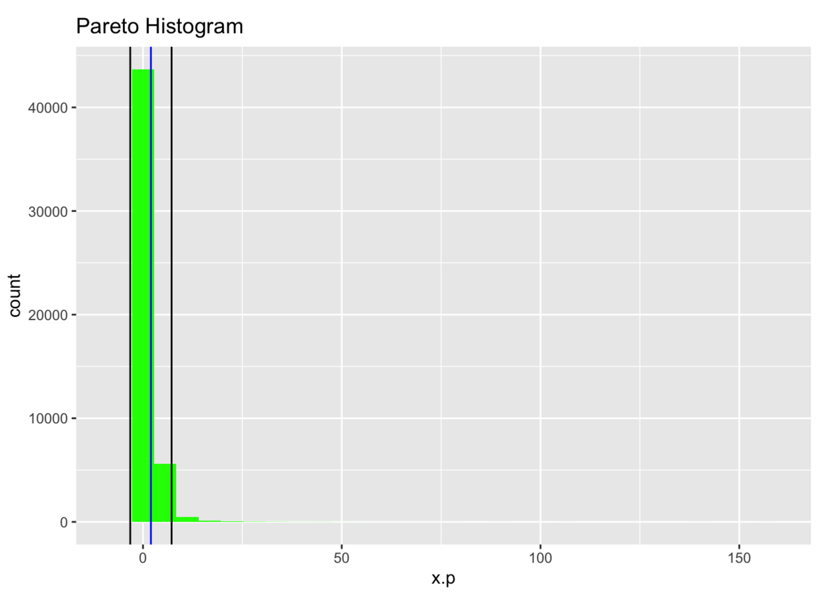
```
## [1] 49036
```

The mean of pareto observations after removing outliers also decreased.

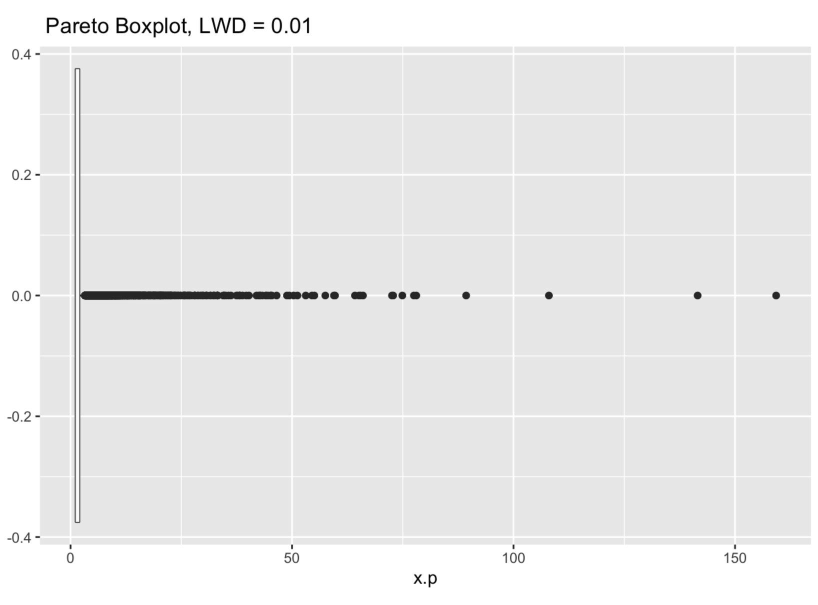
```
P = ggplot(data = Data) + aes( x = x.p) + geom_histogram(fill = "green") + ggtitle("Pareto Histogram") +geom_vlin
e(aes(xintercept=mean(Data$x.p)), color="blue") + geom_vline(xintercept= pmean - 2* psd) + geom_vline(xintercept
 = pmean + 2*psd)
P
```

```
## Warning: Use of `Data$x.p` is discouraged. Use `x.p` instead.
```

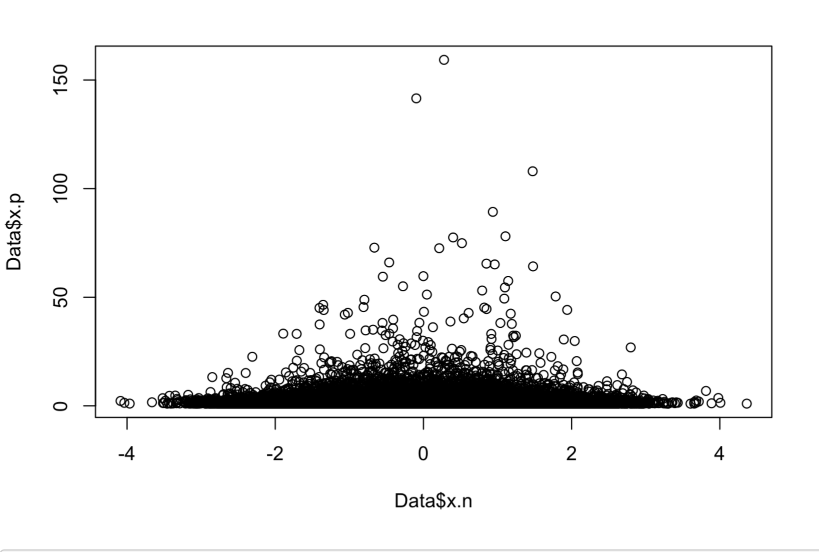
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



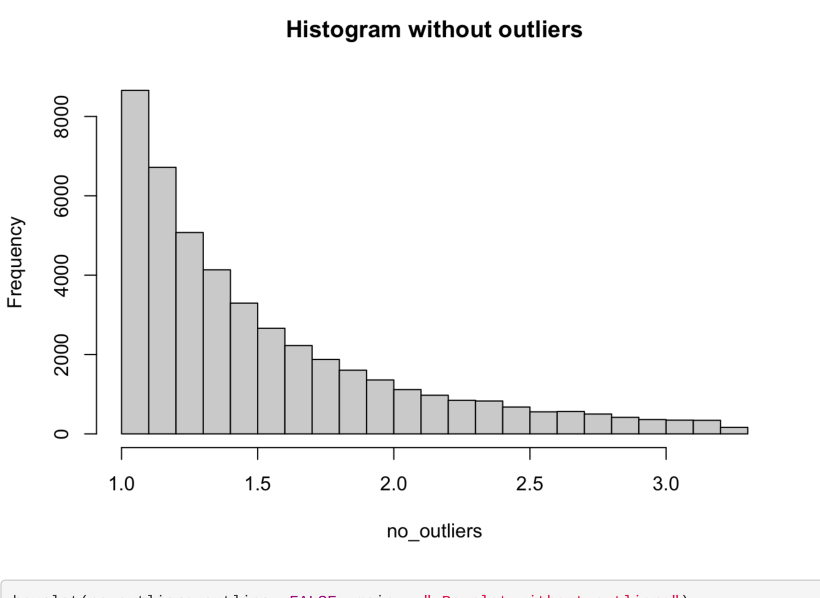

```
C = ggplot(data = Data) + aes(x = x.p) + geom_boxplot(fill = "red") + ggtitle(" Pareto Boxplot, LWD = 0.01") +geo
m_boxplot(lwd=0.01)
C
```




```
plot(Data$x.n,Data$x.p)
```




```
hist(no_outliers, main = "Histogram without outliers")
```




```
boxplot(no_outliers,outline =FALSE, main = " Boxplot without outliers")
```