Assignment 1

Vu The Doan (12918687), Aljer Lee Zhen Yee (12563412)

6/6/2021

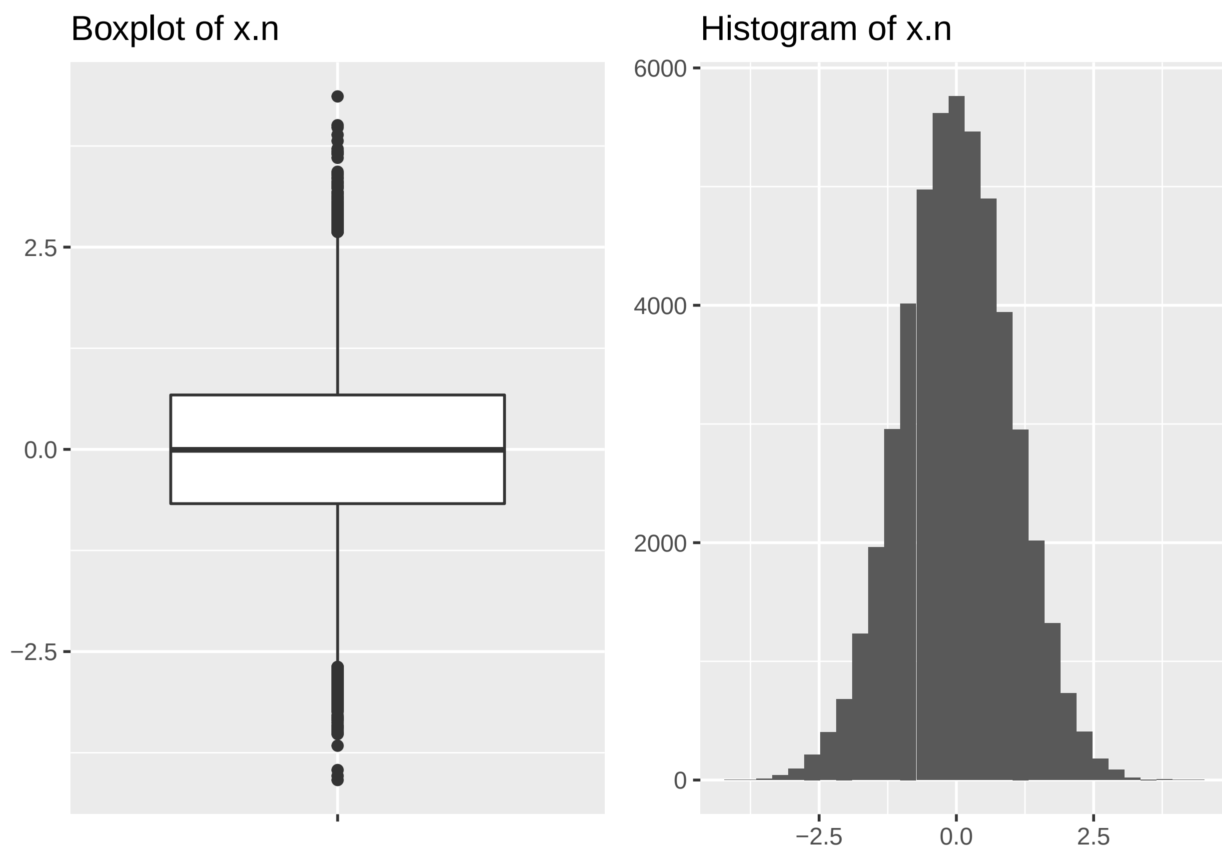
Question 1

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

1. The histogram and boxplot of x.n:

```
library(ggplot2)
library(gridExtra)
P1 <- ggplot(data = Data) +
  geom_boxplot(mapping = aes(x="", y=x.n)) +
  labs(x=NULL, y=NULL) +
  ggtitle("Boxplot of x.n")
P2 <- ggplot(data = Data) +
  geom_histogram(mapping = aes(x=x.n)) +
  labs(x=NULL, y=NULL) +
  ggtitle("Histogram of x.n")
grid.arrange(P1, P2, ncol = 2)
```

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



2. Mean and standard deviation of x.n:

```
attach(Data)
mean(x.n); sd(x.n)
```

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

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

3. The mean is approximately 0 and the standard deviation is approximately 1. Moreover, the histogram and boxplot is symmetric. This corresponds with the assumptions that x.n is a sample of size 50,000 of a normal distribution.

4. Mean and standard deviation of x.p:

```
mean(x.p); sd(x.p)
```

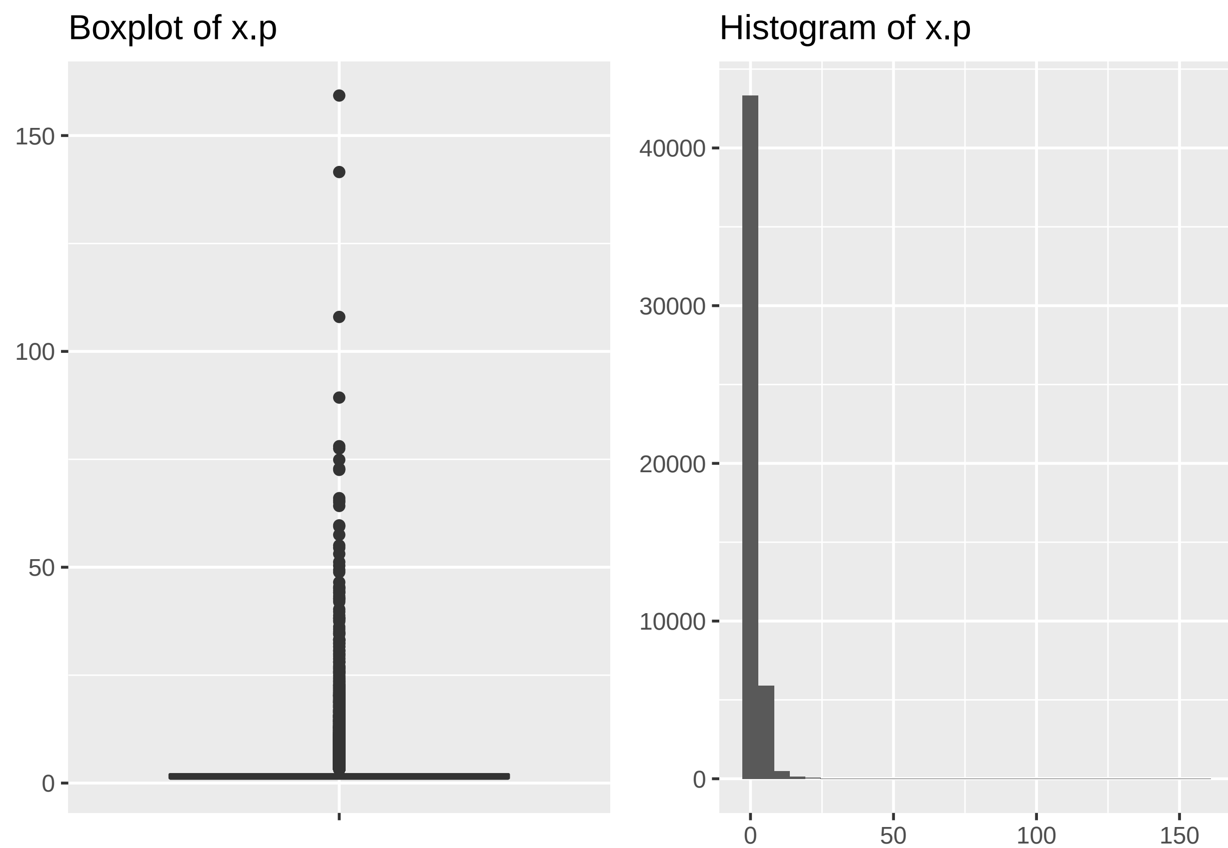
```
## [1] 1.993904
```

```
## [1] 2.601173
```

Histogram and boxplot of x.p:

```
P3 <- ggplot(data = Data) +
  geom_boxplot(mapping = aes(x="", y=x.p)) +
  labs(x=NULL, y=NULL) +
  ggtitle("Boxplot of x.p")
P4 <- ggplot(data = Data) +
  geom_histogram(mapping = aes(x=x.p)) +
  labs(x=NULL, y=NULL) +
  ggtitle("Histogram of x.p")
grid.arrange(P3, P4, ncol=2)
```

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



Standardizing x.p:

```
Data.Z <- (x.p-mean(x.p))/sd(x.p)
mean(Data.Z); sd(Data.Z)
```

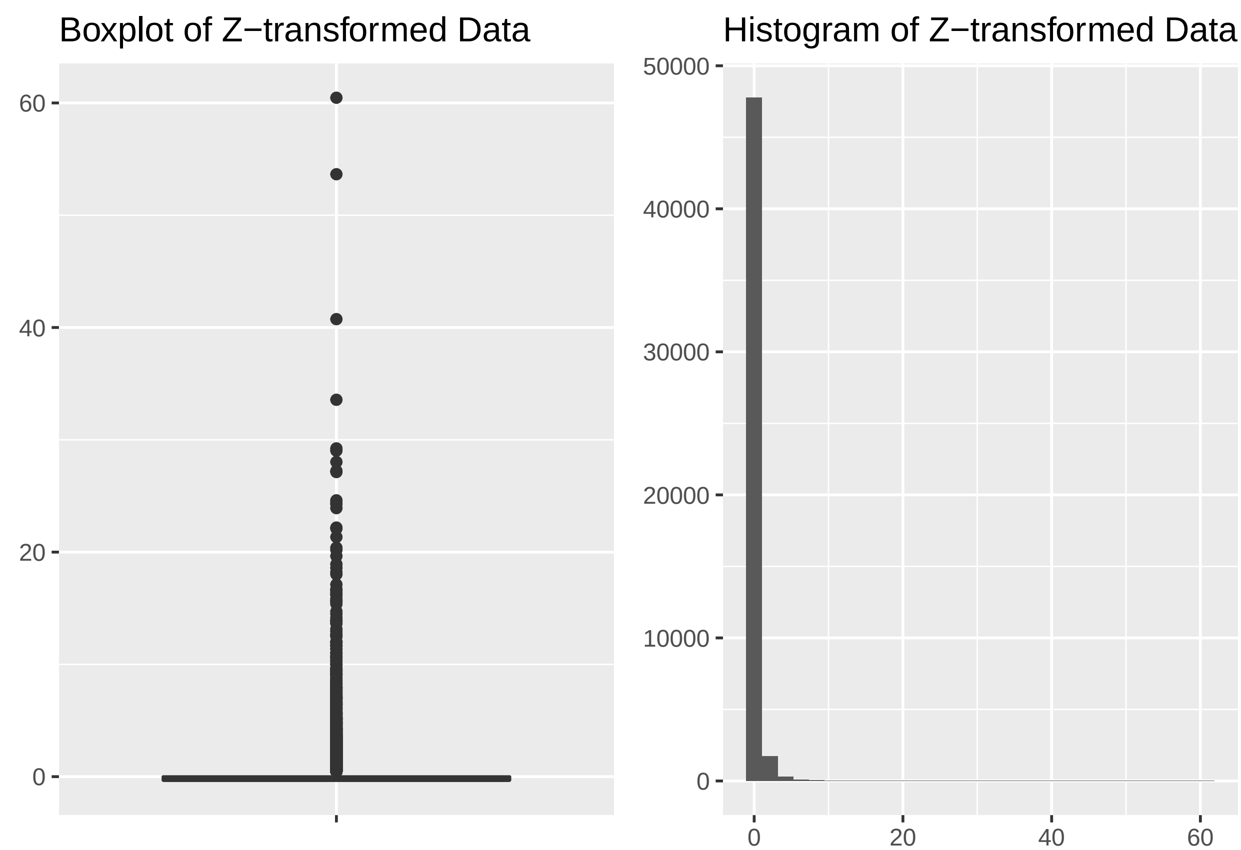
```
## [1] 3.557325e-17
```

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

```
P5 <- ggplot() +
  geom_boxplot(mapping=aes(x="", y=Data.Z)) +
  labs(x=NULL, y=NULL) +
```

```r
  ggtitle("Boxplot of Z-transformed Data")
P6 <- ggplot() +
  geom_histogram(mapping=aes(x=Data.Z)) +
  labs(x=NULL, y=NULL) +
  ggtitle("Histogram of Z-transformed Data")
grid.arrange(P5,P6,ncol=2)
```

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



To be able to compare two data sets that have different distributions, we should standardise the distribution to standard normal distribution ($X \sim N(0, 1)$). A pareto distribution which is commonly known as the 80/20 rule, has a right skew, which can be observed on a graph, where there is a tall "head" to the left and a long tail extending to the right, as it reflects situations in which there are a few items that are very common and a large number that are very rare. For example, 80% of wealth is in the hands of 20% of people.

Application of the Central Limit Theorem (CLT) enable us to approximate pareto distribution to standard normal distribution if n is large enough. CLT states that if we take random sample of a certain distribution and then average it, eventually, for n big enough, the distribution of the sample will eventually be close to normal distribution. "Eventually", could mean an egregiously large n, one greater than the number of samples we can possibly hope to take, which would render the CLT inapplicable. The CLT is about the destination but does not indicate how fast we get there, however, result like the Berry-Esseen Theorem which do bound the rate. In the case of Berry-Esseen, to bounds the largest distance between distribution function of the standardised mean and the standard normal CDF in terms of the third absolute moment ($E(|X|^3)$). So, in the case of the Pareto, if $\alpha > 3$, we can at least get some bound on just how bad the approximation might be at some n, and how quickly we get there. We look at the speed of convergence of the sample means. Based on the data set provided, $\alpha = 2$, hence, the variance does not exist, which implies that the data cannot be transformed, therefore, mean and standard deviation cannot be used to summary the data.

Based on the histogram and boxplot of x.p, it can be observed that the histogram has a right skew, therefore, there are outliers present, which implies that mean is not a good measure of central tendency.

Question 2

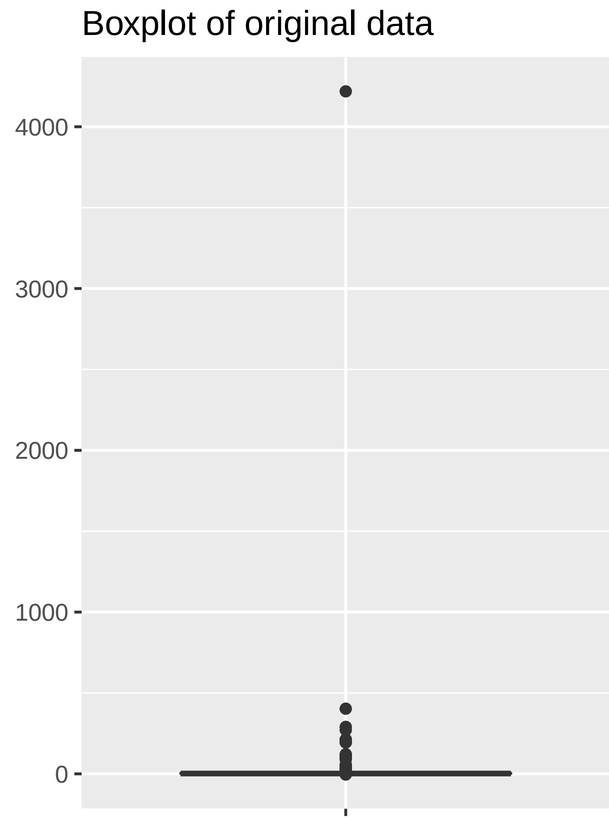
1. Histogram and boxplot of data, unfiltered:

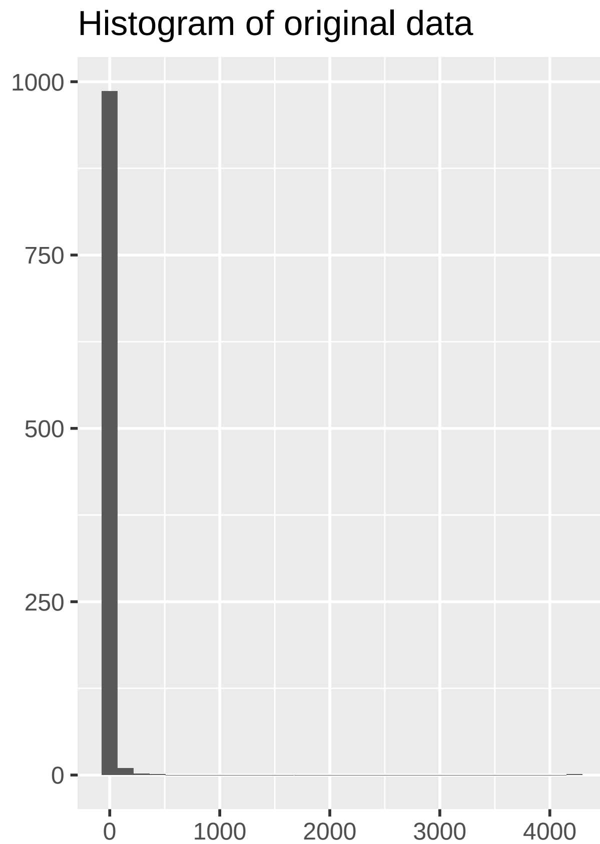
```
Data2 <- read.table("DataAssignment1.txt", sep=","); head(Data2)
```

```
##         V1
## 1 3.894559
## 2 3.516936
## 3 4.568742
## 4 1.529798
## 5 1.368253
## 6 1.477181
```

```
P1 <- ggplot(data = Data2) +
  geom_boxplot(mapping = aes(x="", y=Data2[,1])) +
  ggtitle("Boxplot of original data") +
  labs(x=NULL, y=NULL)
P2 <- ggplot(data = Data2) +
  geom_histogram(mapping = aes(x=Data2[,1])) +
  ggtitle("Histogram of original data") +
  labs(x=NULL, y=NULL)
grid.arrange(P1, P2, ncol=2)
```

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





The summary of Data2 shows that there exists negative values. Since the negative values will produce NAs while being transformed into log scale, they should be removed prior to this operation.

```
summary(Data2)
```

```
##        V1
##  Min.   :  -4.369
##  1st Qu.:   1.332
##  Median :   2.022
##  Mean   :  10.763
##  3rd Qu.:   4.078
##  Max.   :4218.714
```

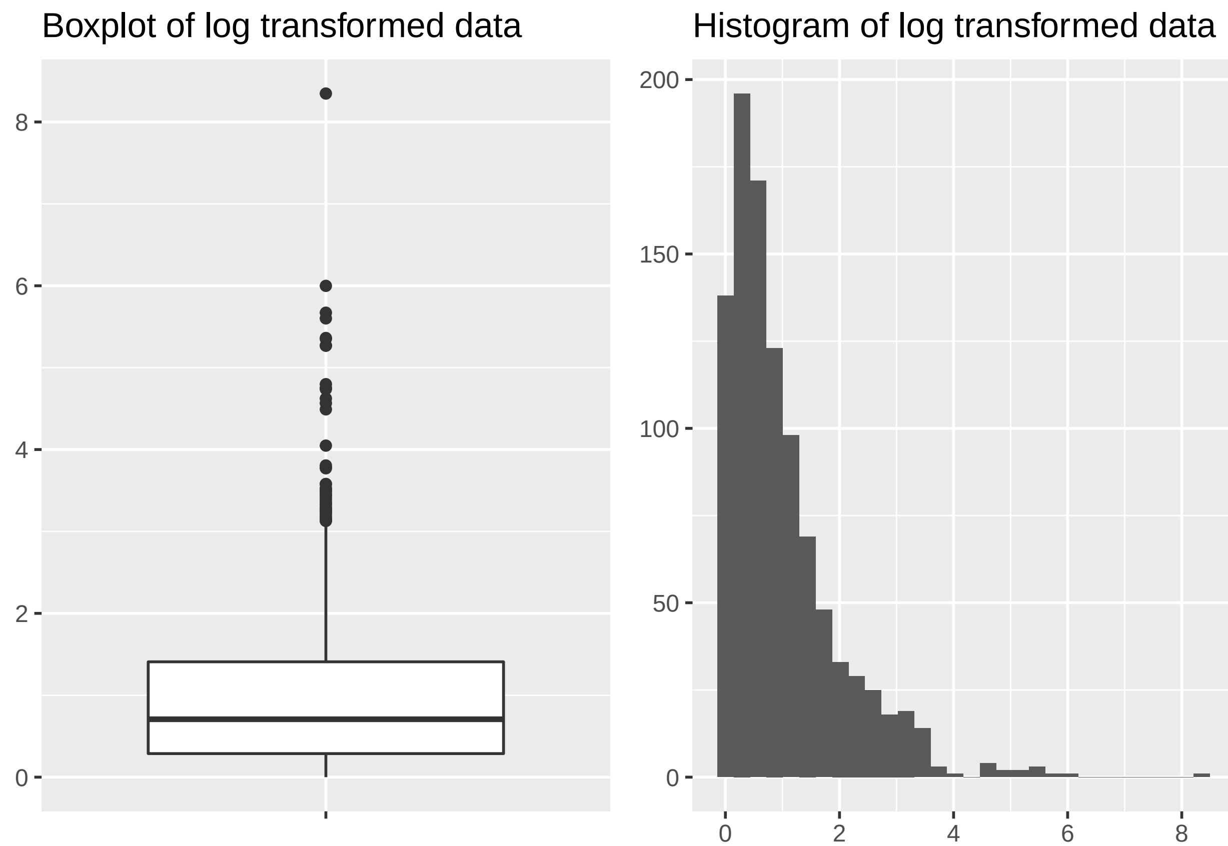
```
Data2_filtered <- Data2[-which(Data2 < 0), 1]
Data2_log <- log(Data2_filtered)
```

Histogram and boxplot of transformed data:

```
P3 <- ggplot() +
  geom_boxplot(mapping = aes(x="", y=Data2_log)) +
  ggtitle("Boxplot of log transformed data") +
  labs(x=NULL, y=NULL)
P4 <- ggplot() +
  geom_histogram(mapping = aes(x=Data2_log)) +
  ggtitle("Histogram of log transformed data") +
```

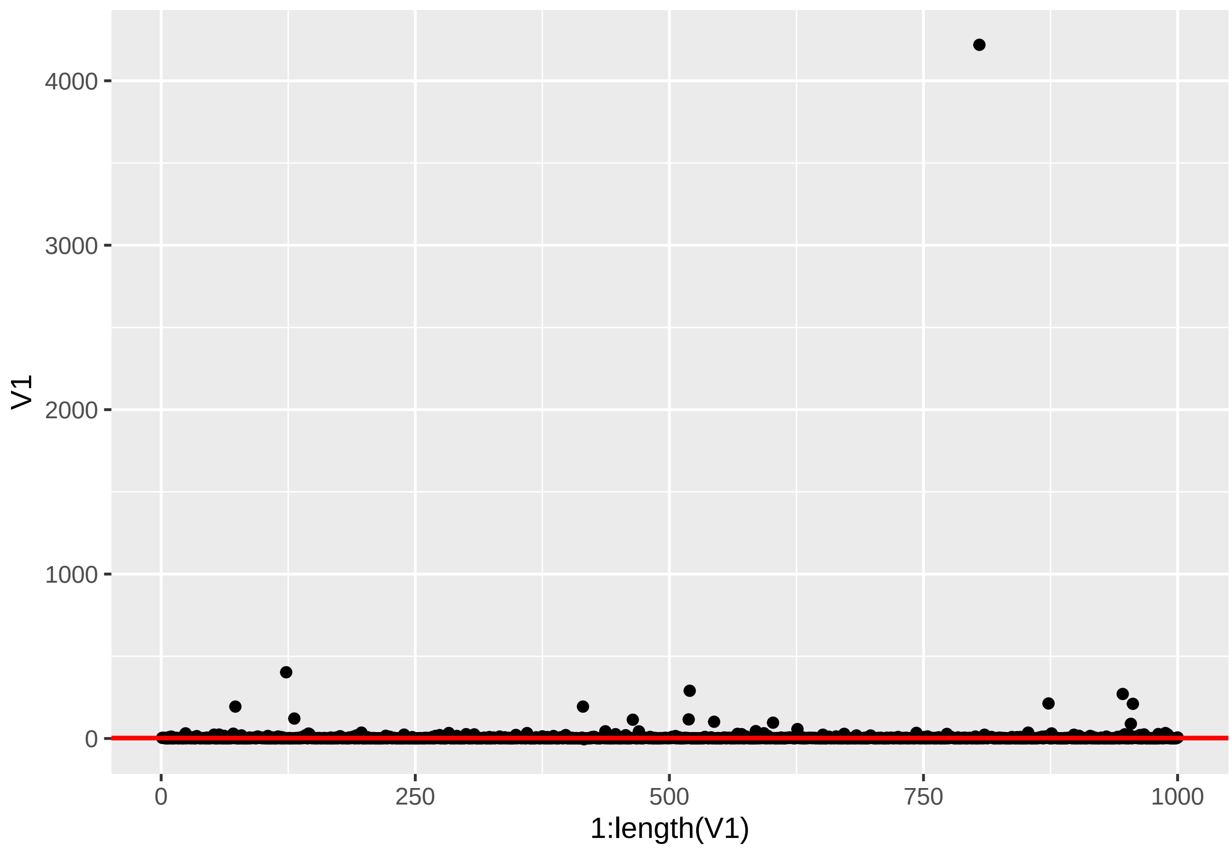
```
  labs(x=NULL, y=NULL)
grid.arrange(P3, P4, ncol=2)
```

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



2. The scatter plot of the original data, with the lower inner fence ($Q1 - 1.5IQR$) and upper inner fence ($Q3 + 1.5IQR$) indicates the outliers:

```
V1 <- Data2$V1
Q1 <- quantile(V1, 0.25); Q3 <- quantile(V1, 0.75); IQR <- Q3 - Q1
ggplot() +
  geom_point(mapping = aes(x=1:length(V1), y=V1)) +
  geom_hline(yintercept = Q1-1.5*IQR, col="red") +
  geom_hline(yintercept = Q3+1.5*IQR, col="red")
```



Here we have a closer look at the outliers:

```
sort(V1[-which(Q1-1.5*IQR < V1 & V1 < Q3 + 1.5*IQR)])
```

```
##   [1]   -4.368772    8.230880    8.380781    8.580403    8.616160    8.656447
##   [7]    8.660911    8.670875    8.694291    8.722244    8.882069    8.906396
##  [13]    8.958837    9.084706    9.171233    9.178223    9.278082    9.292375
##  [19]    9.345212    9.556854    9.696317    9.703077    9.740697    9.824522
##  [25]    9.900632   10.115578   10.263842   10.324007   10.683522   10.757611
##  [31]   11.009740   11.065580   11.116811   11.170467   11.209278   11.619676
##  [37]   11.759001   11.839489   11.840687   11.860067   11.931165   12.158543
##  [43]   12.243893   12.276561   12.497763   12.504606   12.660998   12.719420
##  [49]   12.789380   13.046810   13.078374   13.177983   14.007861   14.136520
##  [55]   14.539494   14.680083   14.880039   14.888120   15.032872   15.072993
##  [61]   15.485777   15.516938   15.526205   15.620091   15.643280   15.855853
##  [67]   15.891467   16.010911   16.400050   16.440444   18.217753   18.400173
##  [73]   18.661672   18.800813   18.979627   19.368565   19.765723   19.913558
##  [79]   20.752573   20.906709   21.009581   21.727357   22.834988   22.930490
##  [85]   23.011650   23.386853   23.547591   23.982069   24.684251   25.152744
##  [91]   25.437938   25.456063   25.768266   26.244951   26.328567   26.435465
##  [97]   26.997615   27.788611   28.268129   28.342793   29.264660   30.105949
## [103]   30.962708   31.310288   31.964917   32.916174   33.211928   33.276831
## [109]   34.141647   35.820794   35.872901   43.444971   43.852069   44.947407
## [115]   57.299963   89.201094   96.196130  101.842394  114.156005  116.088929
## [121]  121.319025  194.071449  194.174139  210.964340  213.208869  270.859804
## [127]  290.449750  402.943895 4218.714040
```

It can be seen that the values -4.368772 and 4218.714040 are the outliers that should be deleted from the data. Further analysis shows that that figures that are above 100 are outliers that should be maintained in the dataset, because it follows a pattern, even though they are significantly higher than the rest of the claim, it makes up the 1% of the claims, which insurance company would have devised such risk. Hence, 402.943895 is the extreme value, which is the highest claim.

```
quantile(V1, 0.99)
```

```
##      99%
## 114.1753
```

The cleaned data is formulated as follows:

```
V1_cleaned <- V1[which(V1 > 0 & V1 < 4000)]
```

3. Mean and median of the cleaned data:

```
mean(V1_cleaned); median(V1_cleaned)
```

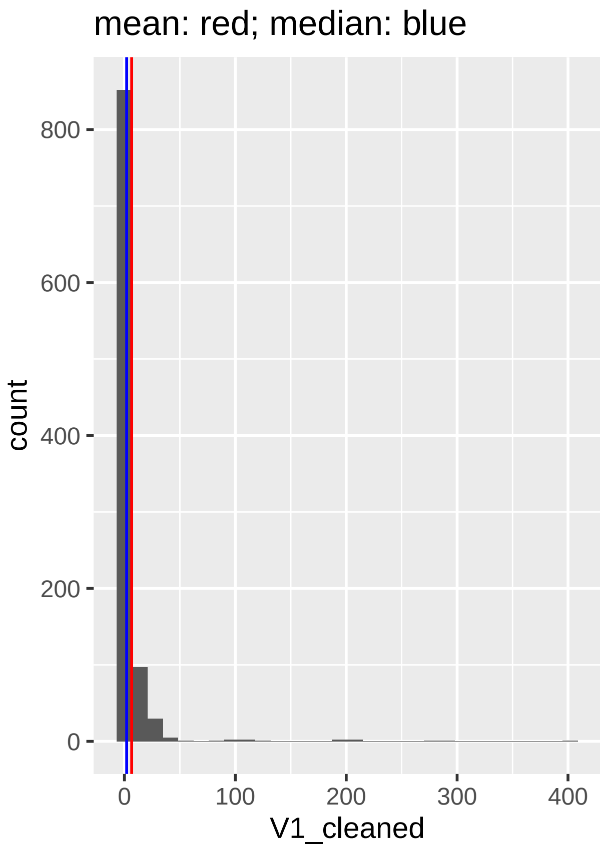
```
## [1] 6.56196
```

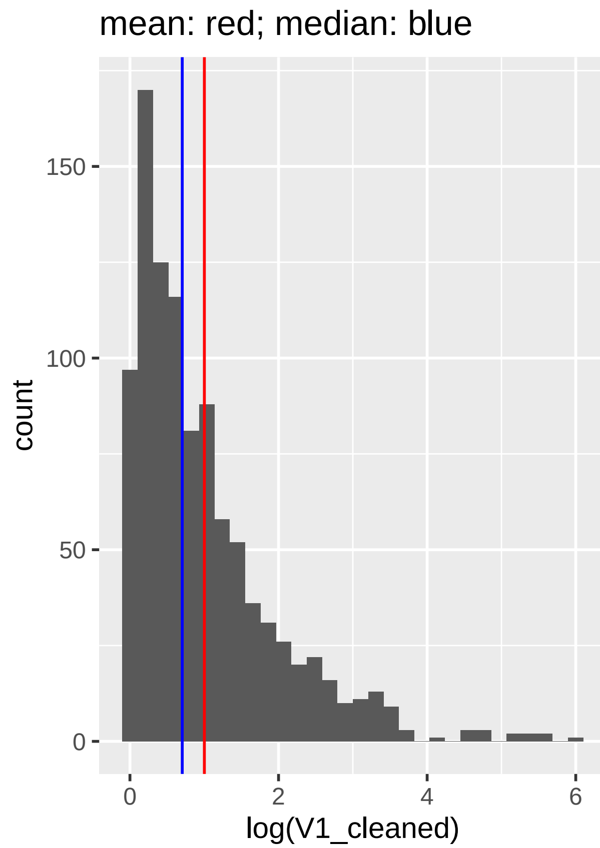
```
## [1] 2.022293
```

The median is lower than the mean, which indicates that the distribution is skewed to the right (positively skewed).

```
P6 <- ggplot() +
  geom_histogram(mapping = aes(x=V1_cleaned)) +
  geom_vline(xintercept = mean(V1_cleaned), col = "red") +
  geom_vline(xintercept = median(V1_cleaned), col = "blue") +
  ggtitle("mean: red; median: blue")
P7 <- ggplot() +
  geom_histogram(mapping = aes(x=log(V1_cleaned))) +
  geom_vline(xintercept = mean(log(V1_cleaned)), col = "red") +
  geom_vline(xintercept = median(log(V1_cleaned)), col = "blue") +
  ggtitle("mean: red; median: blue")
grid.arrange(P6, P7, ncol=2)
```

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





Based on the graphs above, the right skew can be observed. The mean does not deviate too far apart from the median, even though taking into account the outliers in the dataset. Furthermore, a log transformation is done to reduce the skewness and indicates further that it's skewed right, this is due to the majority of the claim of one to two figures in value.

4.

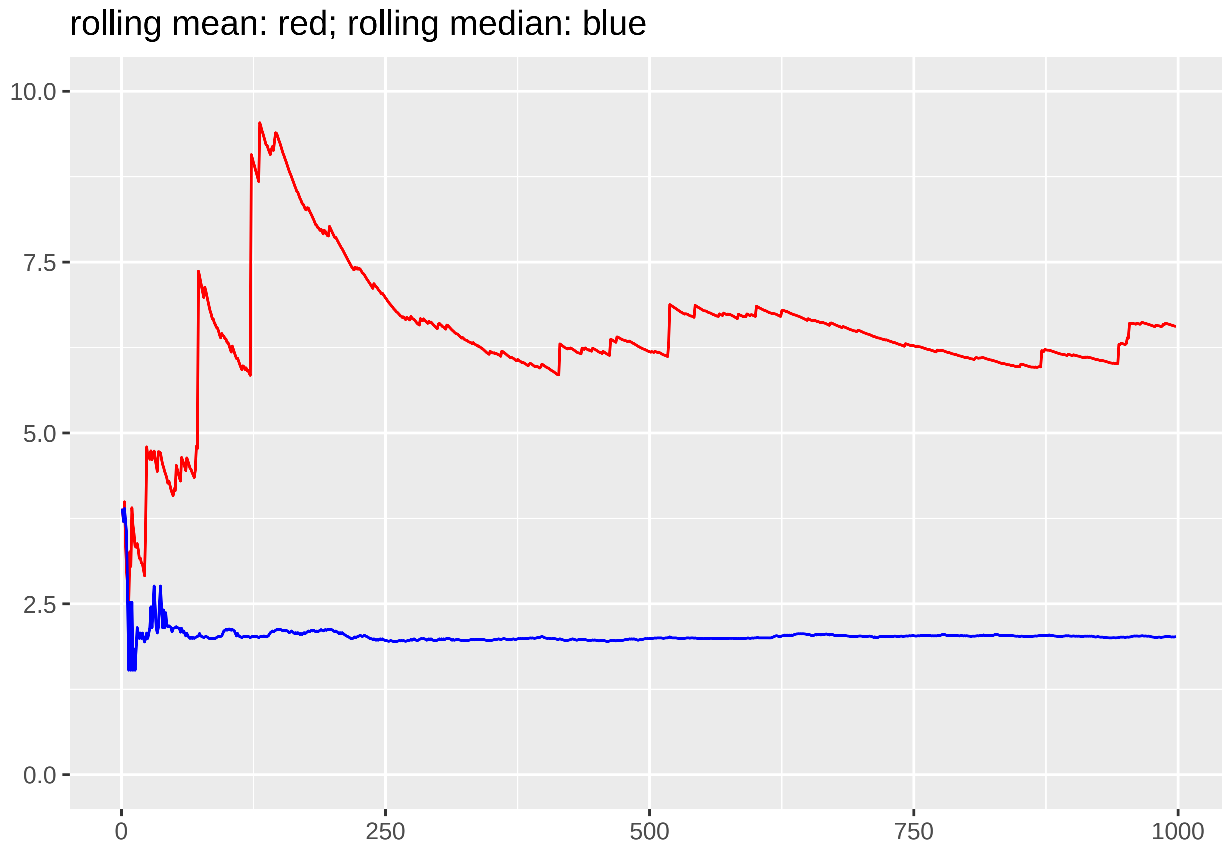
```
rolling_mean <- function(x) {
  cumsum(head(V1_cleaned, n = x))[x]/x
}
rolling_median <- function(x) {
  median(head(V1_cleaned, n = x))
}
rmean <- sapply(1:length(V1_cleaned), rolling_mean); head(rmean)
```

```
## [1] 3.894559 3.705747 3.993412 3.377509 2.975658 2.725912
```

```
rmedian <- sapply(1:length(V1_cleaned), rolling_median); head(rmedian)
```

```
## [1] 3.894559 3.705747 3.894559 3.705747 3.516936 2.523367
```

```
ggplot() +
  geom_line(mapping = aes(x=1:length(rmean), y=rmean), col = "red") +
  geom_line(mapping = aes(x=1:length(rmedian), y=rmedian), col = "blue") +
  ylim(0, 10) +
  labs(y=NULL, x=NULL, title = "rolling mean: red; rolling median: blue")
```



Mean is sensitive to the outliers, hence the fluctuation in values, and based on the graph from subquestion 3 of question 2, the mean does not deviate far from the median, hence both can be used to evaluate central tendency. However, the median is preferred due to its insensitivity to outliers and the right skewness of the histogram.