# Outlier Detection

**Assignment:**

Apply outlier detection techniques to find anomalies in the data set of the number of earthquakes per year with magnitude 7.0 or greater 1900-1998. Data source: https://datamarket.com/data/set/22p8/number-of-earthquakes-per-year-magnitude-70-or-greater-1900-1998#!ds=22p8&display=line

First, I prepared the environment, loaded required libraries and the dataset using the following code:

```
> ### Outlier Detection

> ###Dataset source: https://datamarket.com/data/set/22p8/number-of-earthquakes-per-year-magnitud
e-70-or-greater-1900-1998#!ds=22p8&display=line
>
> rm(list=ls()) #Clear the environment
> setwd("YOUR_PATH") #Set working directory for the assignment
> getwd() #Check working directory
[1] "YOUR_PATH"
>
> #Load libraries
> library(openxlsx)
> library(tseries)
> library(tsoutliers)
> library(forecast)
> library(outliers)
> library(AnomalyDetection)
> library(devtools)
> library(ggplot2)
>
> ###Load Data - file in .xlsx format
> equakes <- read.xlsx("number-of-earthquakes-per-year-m.xlsx", sheet = 1, startRow = 14, colName
s = FALSE, rowNames = FALSE, rows = c(14:112), detectDates = TRUE, fillMergedCells = FALSE )
```

To verify that the data loaded correctly, I looked at the internal structure of the data frame, and the first few and the last few observations. The data frame contains 99 observations of two variables – data and the number of earthquakes per year with 7+ magnitude. Due to differences in handling the origin date between R and Excel, the first data loaded as December

31, 1899, instead of 1/1/1900 as is describes the number of the strong earthquakes registered in 1900.

So, I added the column names, changed the date, and converted the date to a numeric field (for ease of handling 1900) using the following code:

```
> #add column names
> colnames(equakes) <- c("year", "earthquakes")
> equakes$year <- as.numeric(format.Date(equakes$year, format = "%Y", origin ='1900-01-01'))
>
> #change 1989 to 1900 (as the excel date of 1/1/1990 loaded as 12/1/1899)
> equakes[1,1] <- 1900
```

With the following results:

```
> #Check results
> #Check data
> head(equakes)
  year earthquakes
1 1900          13
2 1901          14
3 1902           8
4 1903          10
5 1904          16
6 1905          26
> tail(equakes)
   year earthquakes
94 1993          16
95 1994          15
96 1995          25
97 1996          22
98 1997          20
99 1998          16
> #check structure of the df
> str(equakes)
'data.frame':      99 obs. of  2 variables:
 $ year       : num  1900 1901 1902 1903 1904 ...
 $ earthquakes: num  13 14 8 10 16 26 32 27 18 32 ...
```

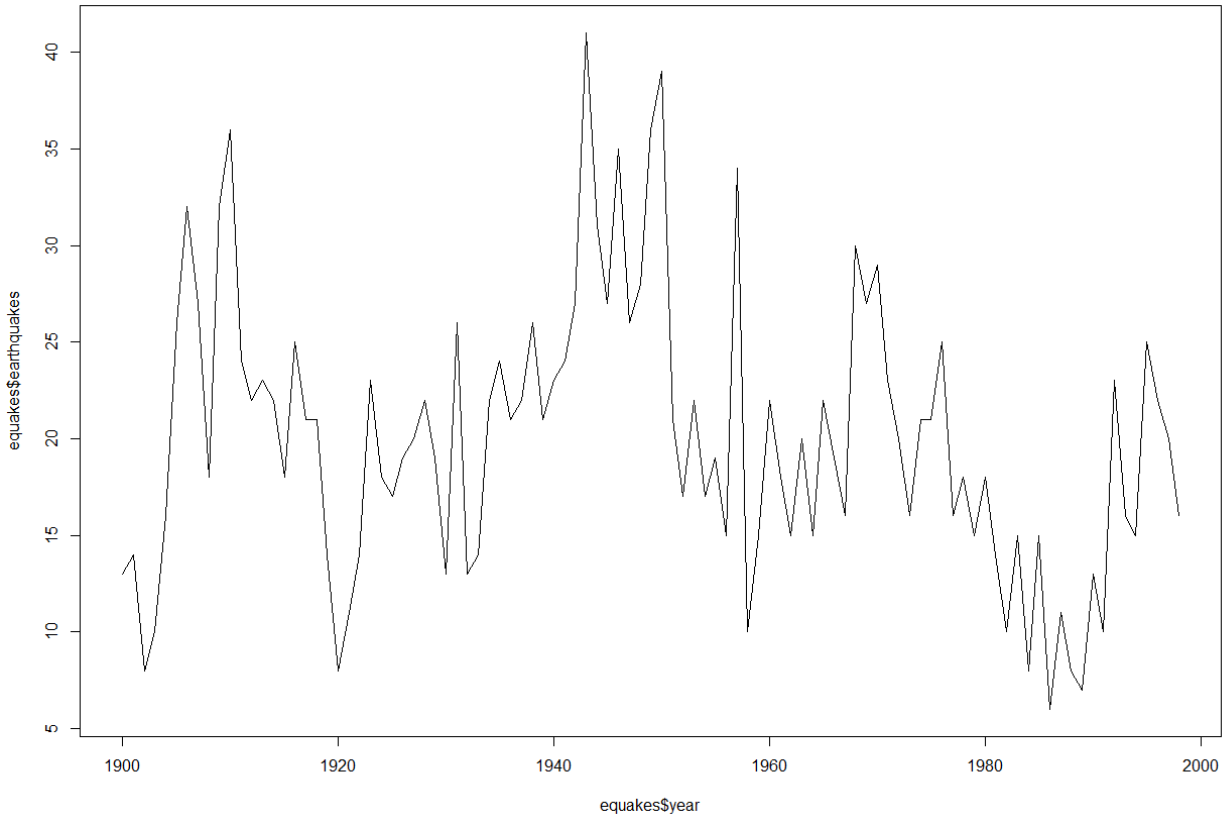
Next, I looked at the summary statistics for the dataset variables. While statistics for the year field do not make much sense as I intentionally kept the date in a simple numeric format, the summary statistics for the earthquakes filed show that all 99 observations fall between 6 and 41 string earthquakes per year. The median (20.0) and the mean (20.02) values are very close, suggesting that the distribution might be close to symmetric.

```
> #####EDA, check for normal distribution
> summary(equakes)
      year       earthquakes
 Min.   :1900   Min.   : 6.00
 1st Qu.:1924   1st Qu.:15.00
```

```
 Median :1949   Median :20.00
 Mean   :1949   Mean   :20.02
 3rd Qu.:1974   3rd Qu.:24.00
 Max.   :1998   Max.   :41.00
```

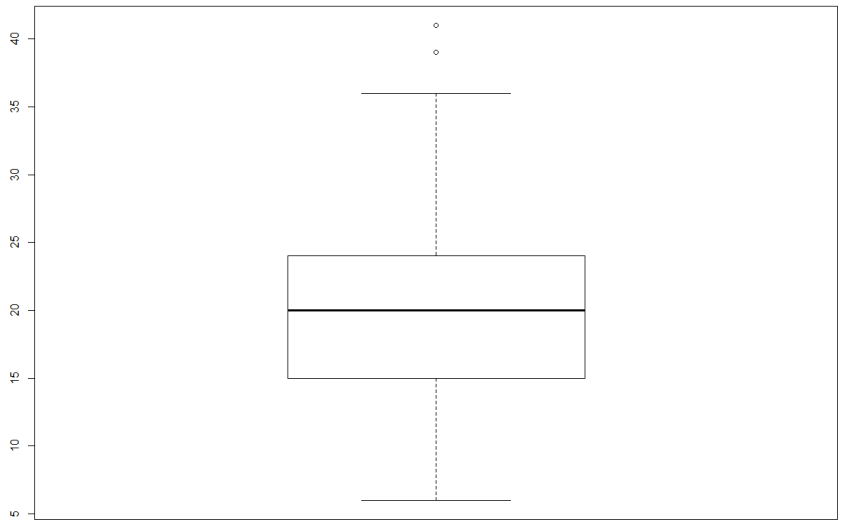
A graph of the observations shown below demonstrates that the number of earthquakes fluctuates significantly and there are no obvious trends in the data:

```
> plot(equakes$earthquakes ~ equakes$year, type='l')
```




Boxplot of the earthquake data shows a nearly symmetrical distribution and two potential outliers just above and below 40 earthquakes per year.

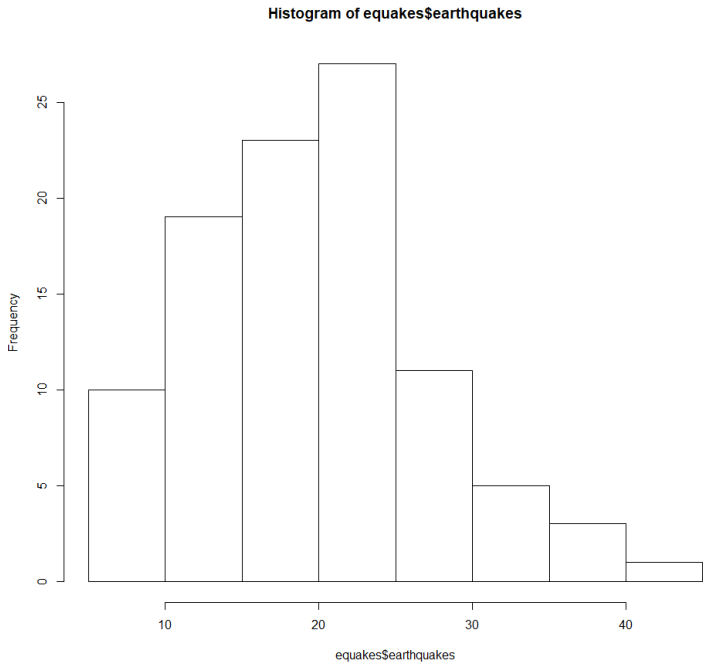
```
> boxplot(equakes$earthquakes) #boxplot
```



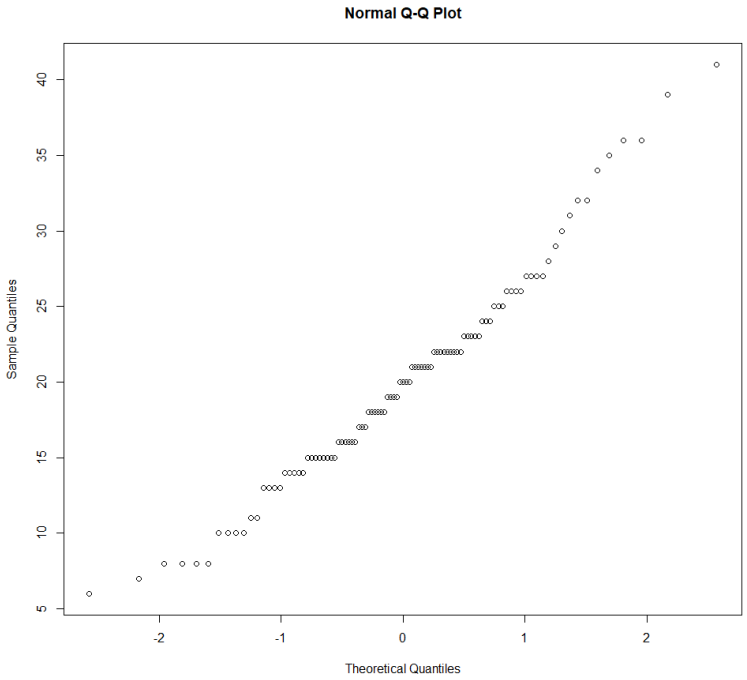

In order to choose statistical tests that can be applied to investigate the potential outliers further, we need to establish whether the earthquake data is normally distributed or not.

In order to visually inspect the normality, I used the histogram and quantile-to-quantile plots.

```
> hist(equakes$earthquakes) #histogram
```



```
> qqnorm(equakes$earthquakes) #normal quantile-quantile plot
```




Both of these graphs who that the data is very close to normal, but in order to have a definitive answer I ran the Shapiro-Wilk normality test with the following hypothesis:

H0: The earthquake data is normally distributed;

Ha: The earthquake data is not normally distributed.

```
> #Shapiro-Wilk normalisty test
> shapiro.test(equakes$earthquakes)

        Shapiro-Wilk normality test

data:  equakes$earthquakes
W = 0.97538, p-value = 0.0601
```

The above results show that the test returned p-value of 0.0601 which is higher than the significance level of 0.05. It means that we do not have enough evidence to reject the null hypothesis. The earthquake data is not significantly different from the normally distributed. It allows us to use parametric statistical tests for detecting potential outliers in the dataset.

First, I used the Grubbs' test. The boxplot suggested that there might be two potential outliers close to the 40 earthquakes per year. Unfortunately, the grubbs.test() with type=20 capable to detect two outliers can work only on small datasets of 3-30 observations. So, I used the Grubbs' test to look for at least one outlier on the higher side of the observations using the following hypothesis:

H0: There are no outliers in the dataset;

Ha: There is an outlier in one tail (max value of 41 is an outlier).

```
> grubbs.test(equakes$earthquakes, type = 10, opposite = FALSE, two.sided = FALSE)

        Grubbs test for one outlier

data:  equakes$earthquakes
G = 2.88850, U = 0.91399, p-value = 0.1594
alternative hypothesis: highest value 41 is an outlier
```

This test returned the p-value of 0.1594, which is higher than the statistical significance level of 0.05. It means that we do not have enough evidence to reject the null hypothesis in favor of the alternative hypothesis. So, according to the Grubbs' test, the maximum value of 41 that was observed in 1943 is not an outlier in the dataset.

I also tan the Chi-Squared test for outliers with the following hypothesis:

H0: There are no outliers in the dataset;

Ha: There is at least one outlier in the dataset (the maximum value of 41 is an outlier).

```
> chisq.out.test(equakes$earthquakes)

        chi-squared test for outlier

data:  equakes$earthquakes
X-squared = 8.3434, p-value = 0.003871
alternative hypothesis: highest value 41 is an outlier
```

As the above results show, the according to the Chi-Squared test for outliers, we have enough evidence to reject the null hypothesis, as the test returned a small p-value of 0.003871,

with is smaller than the statistical significance level of 0.05. It means that the null hypothesis should be rejected in favor of the alternative hypothesis stating that 41 is, in fact, an outlier.

In the two statistical test yielded completely opposite results concerning the maximum value of 41. At the same time, neither of them was concerned with the minimum value in the dataset. These statistical tests also ignore the temporal aspect of the dataset, so I decided to use the time series approach.

I created a time series object and used auto.arima() to automatically chose parameters for the ARIMA model.

```
> #create a time series object
> eqts <- ts(equakes$earthquakes, start=c(1900), end=c(1998), frequency = 1)
> #automatically choose model parameteres
> eq_fit <- auto.arima(eqts)
> eq_fit
Series: eqts
ARIMA(1,0,1) with non-zero mean

Coefficients:
         ar1      ma1     mean
      0.8308  -0.4371  19.6675
s.e.  0.0863   0.1369   1.9010

sigma^2 estimated as 35.75:  log likelihood=-316.23
AIC=640.46   AICc=640.88   BIC=650.84
```

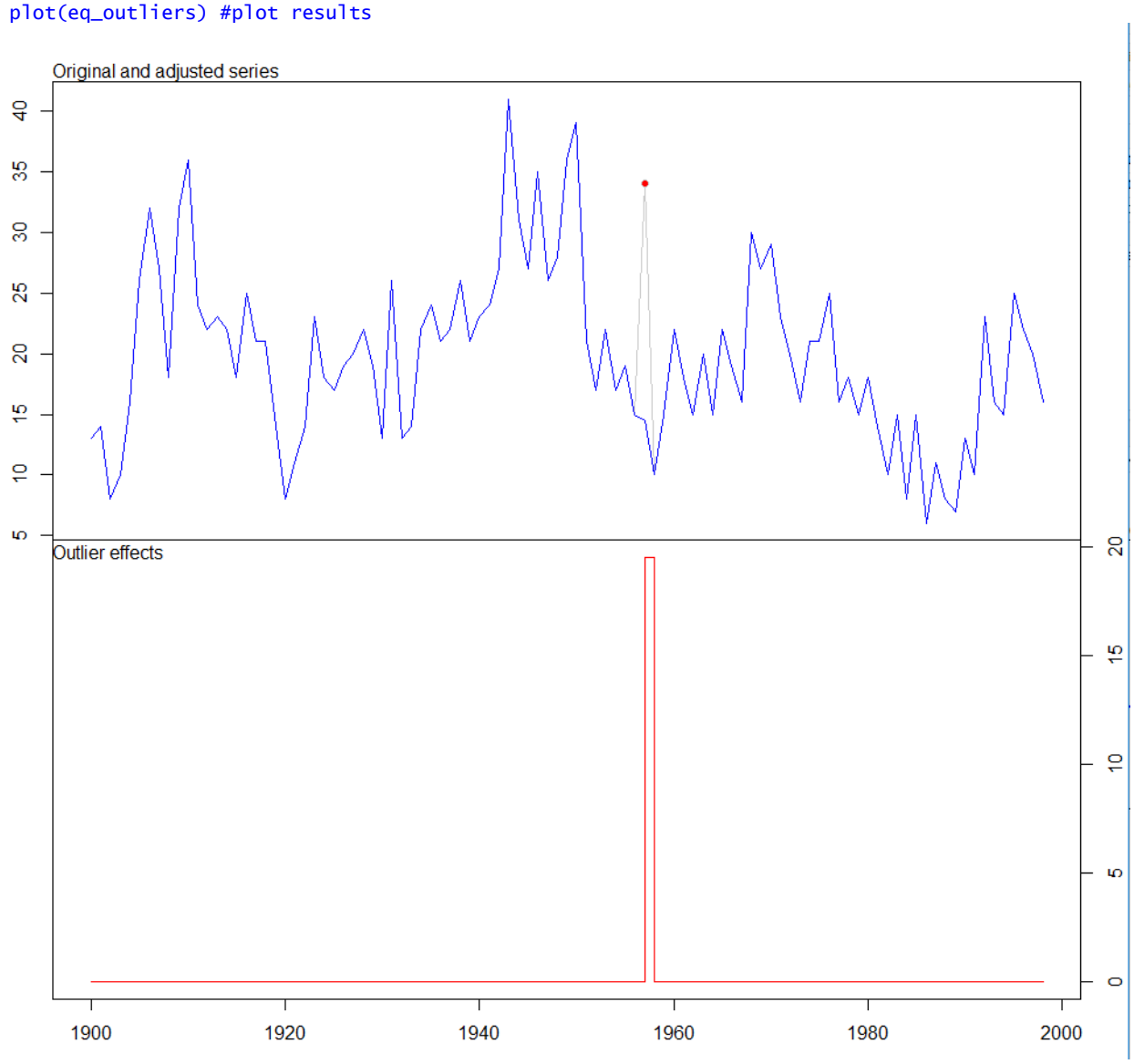
It resulted in the ARIMA(1, 0, 1) model as the best fit. Next, I used these parameters with the tso() function from the tsoutliers package to automatically detect the outliers.

```
> #use ARIMA (1,0,1) model and the tsp function to highlight the outliers fro
m the dataset
> eq_outliers <- tso(eqts, tsmethod = 'arima', args.tsmethod = list(order = c
(1,0,1)))
> eq_outliers #detailed model info

Call:
list(method = NULL)

Coefficients:
         ar1      ma1  intercept     AO58
      0.7987  -0.3081    19.5121  19.5218
s.e.  0.0986   0.1653     1.8399   4.9570

sigma^2 estimated as 30.14:  log likelihood = -309.34,  aic = 628.68
```

```
Outliers:
  type ind time coefhat tstat
1   AO  58 1957   19.52 3.938
> plot(eq_outliers) #plot results
```

According to the ARIMA model, the number of the earthquakes (34) with the magnitude of 7+ recorded in 1957 is an outlier that has considerable influence on the time series, as demonstrated by the graph below:

```
plot(eq_outliers) #plot results
```



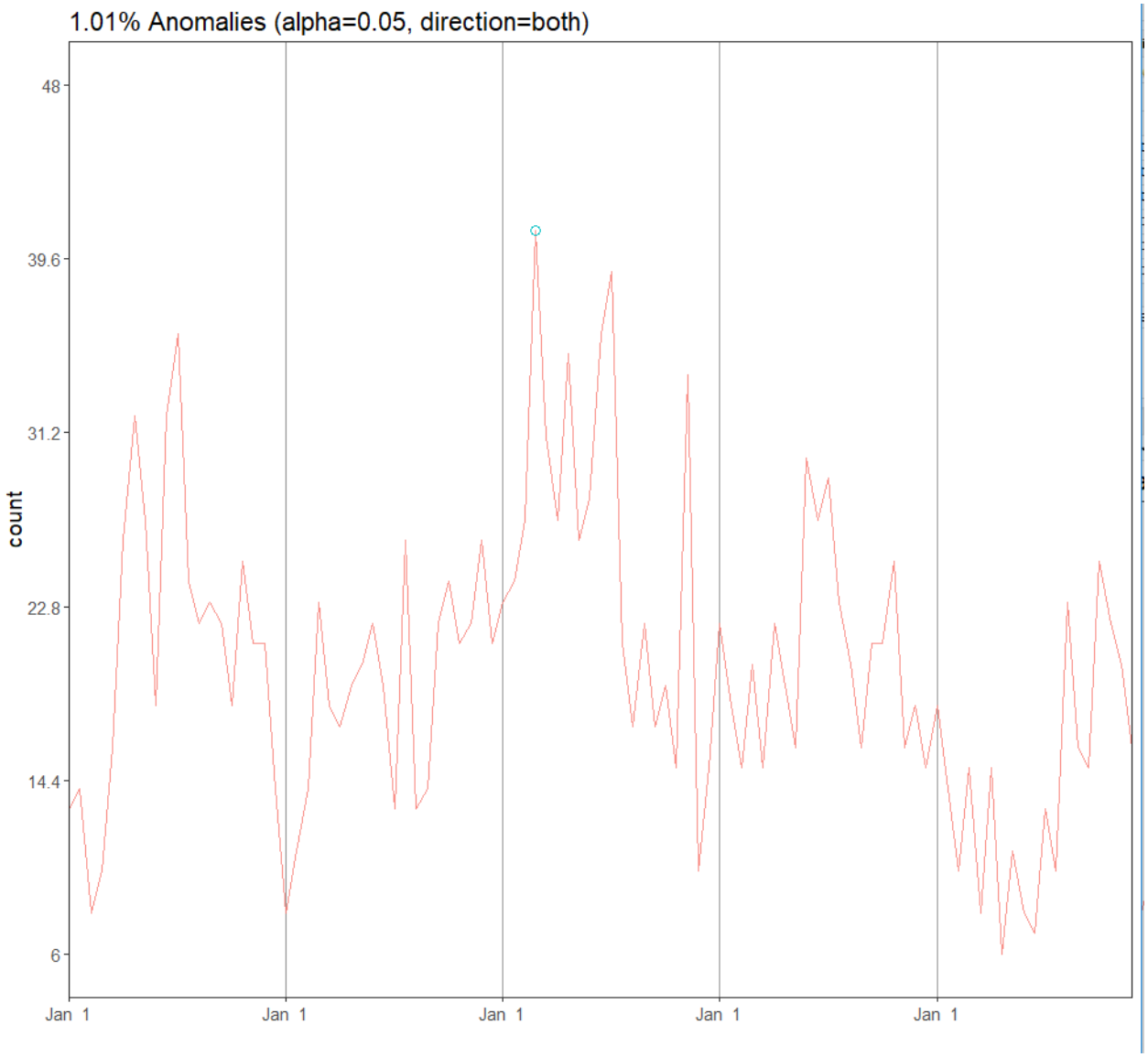
While 34 earthquakes per year are neither a global maximum or global minimum, it represents a local maximum that completely reverses the trend observed in the preceding and following years.

Next, I used the AnomalyDetection package. I converted the data in the format accepted by this package using the following code:

```
> #df dates from 01/01/1900 to 01/01/1998
> years <- as.data.frame(seq(as.Date("1900/1/1"), as.Date("1998/1/1"), "years"))
> colnames(years) <- c('year')
> #convert date to POSIXlt format
> years$year<-as.data.frame(as.POSIXlt(years$year,format = '%Y'))
> #create df to be used with the AnomalyDetection package
> equakes2<- cbind.data.frame(years$year,equakes$earthquakes)
> colnames(equakes2) <- c("year", "earthquakes") #rename columns
```

Next, I used the AnomalyDetectionTS() command to identify the anomalies:

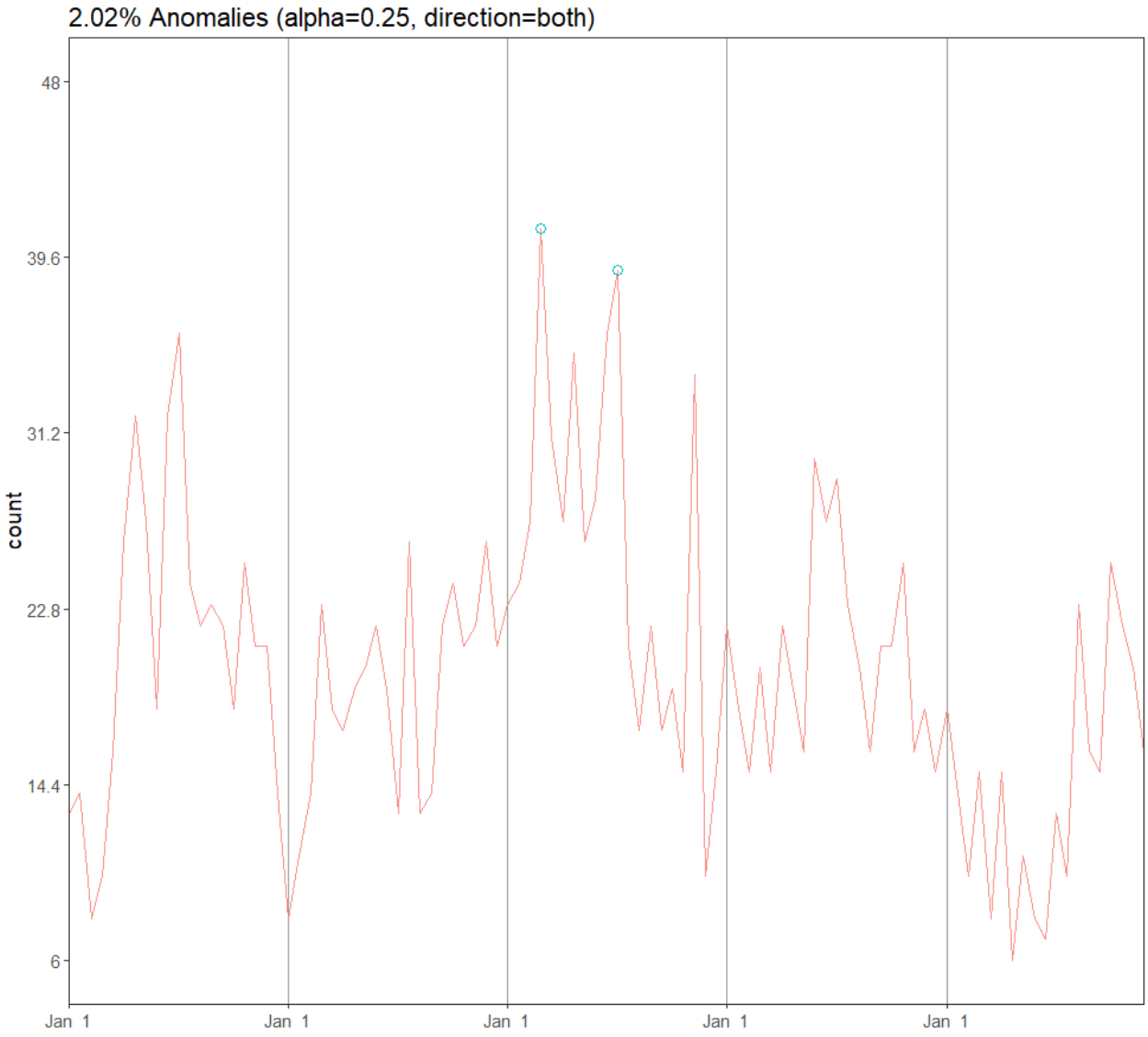
```
> #AnomalyDetection.AnomalyDetectionTs(equakes, plot=TRUE)$plot
> anomalies<- AnomalyDetectionTs(equakes2, direction = 'both', max_anom = 0.1, plot  = TRUE)
> anomalies$plot
> anomalies$anoms
   timestamp anoms
1 1943-01-01    41
```




As the above results and the graph show, the global maximum of 41 earthquakes observed in 1943 once again was identified as an outlier (anomaly). No other anomalies were detected even is the maximum number of anomalies was set to 0.1 in the command used. In order to see if this method can potentially pick up any other anomalies in the dataset, I the alpha parameter, which controls the level of significance for potential anomalies, by setting it to 0.25 (as opposed to 0.05 by default in the previous case).

```
> #alpha determines level of significance for anomalies, default alpha = 0.05
> #decrease to alpha= 0.25
```

```
> anomalies2<- AnomalyDetectionTs(equakes2, direction = 'both', alpha=0.25,max_anom = 0.1, plot
= TRUE)
> anomalies2$plot
> anomalies2$anoms
   timestamp anoms
1 1943-01-01    41
2 1950-01-01    39
```




It resulted in one more anomaly being detected 39 earthquakes observed in 1950 which is a local maximum, the second largest number of the earthquakes in the dataset. These results correspond to the two potential outliers originally suggested by the boxplot.

In conclusion, this exercise demonstrated that the outlier detection problem is not a straightforward process. Its results are highly influenced by the tools used in the analysis and need to take into consideration the nature of data being analyzed. Then earthquake dataset does not contain any data points that, based on a background domain knowledge, would clearly identify some observations as anomalies. The simplest approach used – the boxplot- provided a good first suggestion about possible outliers based on distance. The statistical tests are convenient for identifying possible outliers among global maximums and minimums, however, they used different test statistics and can yield opposite results as above. The AnomalyDetection package, depending on the settings used, can additionally identify local maximums and minimums as potential outliers. This approach can be beneficial for monitoring ongoing processes where global extremes might never be known (e.g., online user behavior, intrusion detection). The time series analysis is the only approach used that was able to identify a potential outlier that had a significant influence on the overall trend in the time series.

Overall, the most beneficial approach to identifying outliers seems to be using a combination of methods to identify potential anomalies that then would need to be further evaluated depending on the domain area and the goal (e.g., finding and eliminating possible technical errors in the dataset before applying other analysis techniques or detecting outliers as the main goal of the analysis).