Week 7 Anomaly Detection

Ryan Schraeder

2022-04-24

In this week's exercise, I'll be diving into outlier detection methods in the provided yearly earthquake data. First and foremost, I'll need to take a look at the data to assess data types, observations, and central tendencies.

Exploratory Analysis

```
data<-read.csv('/Users/rschraeder/Downloads/wk7_eq.csv')
str(data)
```

```
## 'data.frame':    99 obs. of  2 variables:
##  $ year       : int  1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 ...
##  $ earthquakes: int  13 14 8 10 16 26 32 27 18 32 ...
```

As it appears, there are 99 observations with two variables 1. An integer type variable indicating year. 2. An integer type variable indicating number of earthquakes within the given year.

```
sum(is.na(data))
```

```
## [1] 0
```

There are no null values, so imputation won't be necessary.

```
summary(data)
```

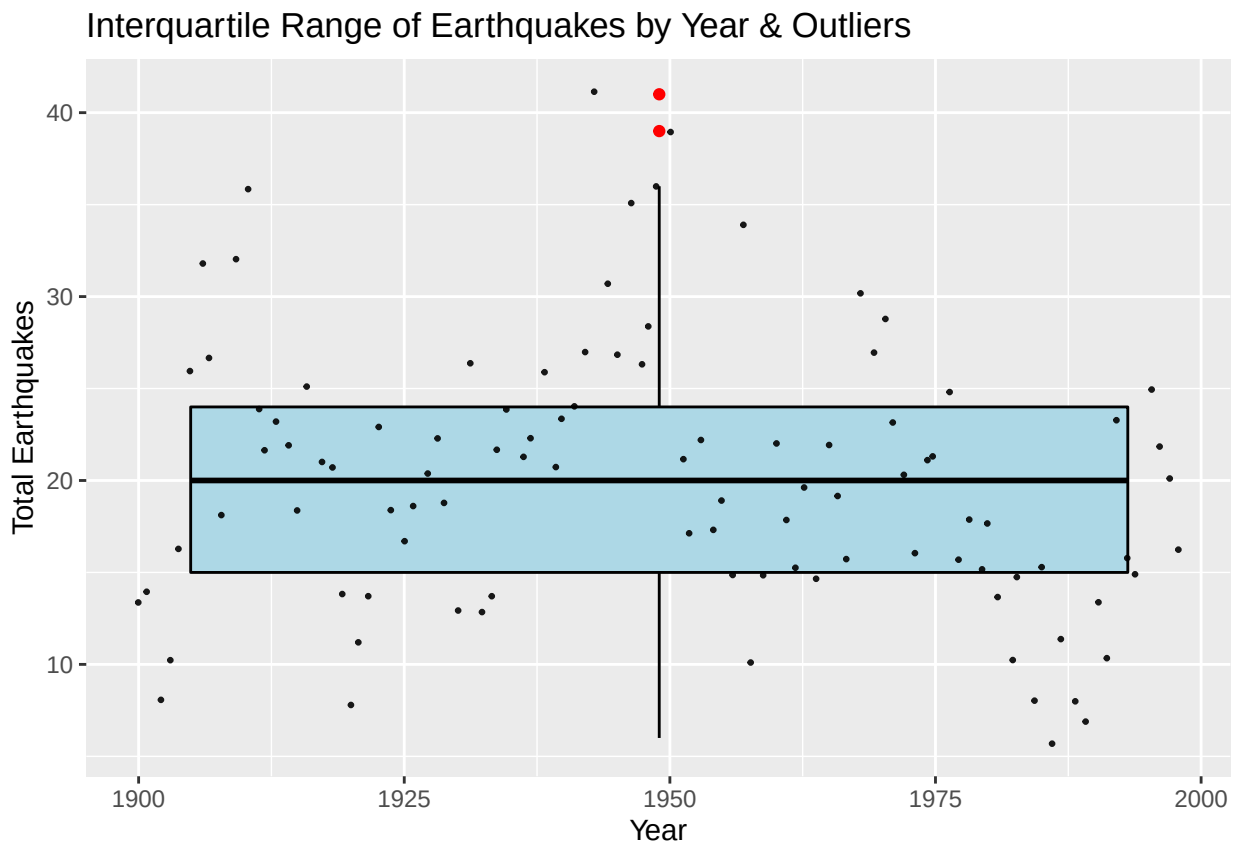
```
##       year       earthquakes
##  Min.   :1900   Min.   : 6.00
##  1st Qu.:1924   1st Qu.:15.00
##  Median :1949   Median :20.00
##  Mean   :1949   Mean   :20.02
##  3rd Qu.:1974   3rd Qu.:24.00
##  Max.   :1998   Max.   :41.00
```

Summarizing the data, our central tendences show that the timeline is about 100 years, the average amount of earthquakes per year is 20, and the most earthquakes witnessed in a given year is 41. Let's continue the exploratory analysis with a boxplot to better visualize these statistics and understand the distribution of our data.

Plots

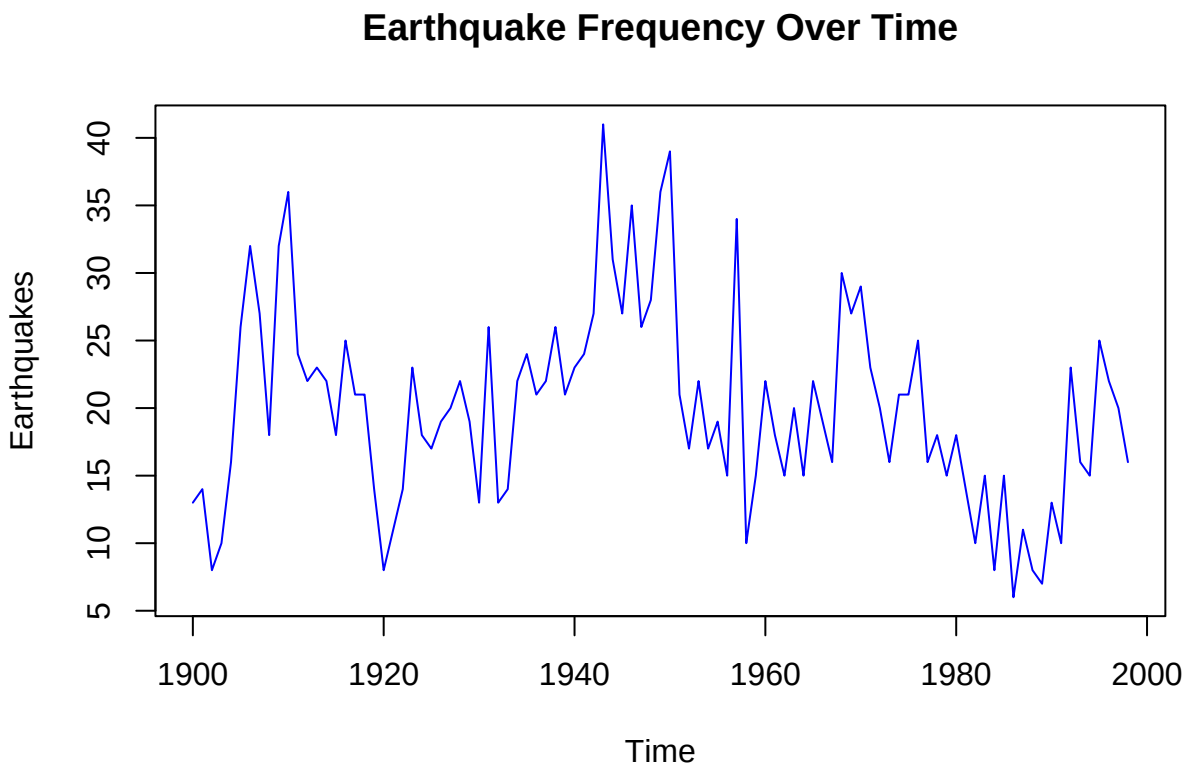
```
## Boxplot
ggplot(data, aes(x = year, y = earthquakes)) +
  geom_boxplot(color = "black", fill = "lightblue", outlier.color = "red") +
  geom_jitter(color='black', size=0.5, alpha=0.9) +
  ggtitle("Interquartile Range of Earthquakes by Year & Outliers") +
  xlab("Year") +
  ylab("Total Earthquakes")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



As we can see, the scatter of our data points suggests an up and down trend of earthquake activity with few outliers in 1949. We can get a better idea of the trend of these earthquakes by converting the data to a time-series object.

```
ts_data <- ts(data$earthquakes, start=1900, end=1998, frequency=1)
plot(ts_data, main="Earthquake Frequency Over Time", ylab="Earthquakes", col="blue")
```



We can see the trend outlined in this line plot, where the volume of earthquakes has many spikes between 1939-1950 and an unexpected spike just before 1960. However, we can tell the data is rather sporadic. This case will be a perfect example of using anomaly detection to understand any years where earthquake volume increased significantly outside of tendencies in our data.

Anomaly Detection Methods

Statistical Methods

Using statistical methods, we can measure significance tests to locate outliers in our data. Several tests can be used for this, and we can compare outliers on both ends of the data to see if there is a consistent pattern in outlier detection.

```
# Sample
sample<- sample(data$earthquakes,30)

#Dixon Test
dixon.test(sample)
```

```
##
##  Dixon test for outliers
##
## data:  sample
## Q = 0.40909, p-value = 0.05506
## alternative hypothesis: highest value 36 is an outlier
```

```
#Grubbs Test
grubbs.test(sample)
```

```
##
##  Grubbs test for one outlier
##
## data:  sample
## G = 2.65909, U = 0.74777, p-value = 0.07025
## alternative hypothesis: highest value 36 is an outlier
```

```
#Chi-Squared Test
chisq.out.test(sample)
```

```
##
##  chi-squared test for outlier
##
## data:  sample
## X-squared = 7.0708, p-value = 0.007835
## alternative hypothesis: highest value 36 is an outlier
```

Using a sample of our data of 30 records, each test can be run. Collectively, the alternative hypothesis of the high end (right tail) of the data being an outlier is proven, given p-values are less than 0.05. The Grubbs test displays proof for the null hypothesis, but we can attest the upper boundary of the data is non-normal.

Testing for the lower boundary may conclude our testing.

```
#Dixon Test
dixon.test(sample, opposite=TRUE)
```

```
##
##  Dixon test for outliers
##
## data:  sample
## Q = 0.23529, p-value = 0.6223
## alternative hypothesis: lowest value 10 is an outlier
```

```
#Grubbs Test
grubbs.test(sample, opposite=TRUE)
```

```
##
##  Grubbs test for one outlier
##
## data:  sample
## G = 1.75387, U = 0.89027, p-value = 1
## alternative hypothesis: lowest value 10 is an outlier
```

```
#Chi-Squared Test
chisq.out.test(sample, opposite=TRUE)
```

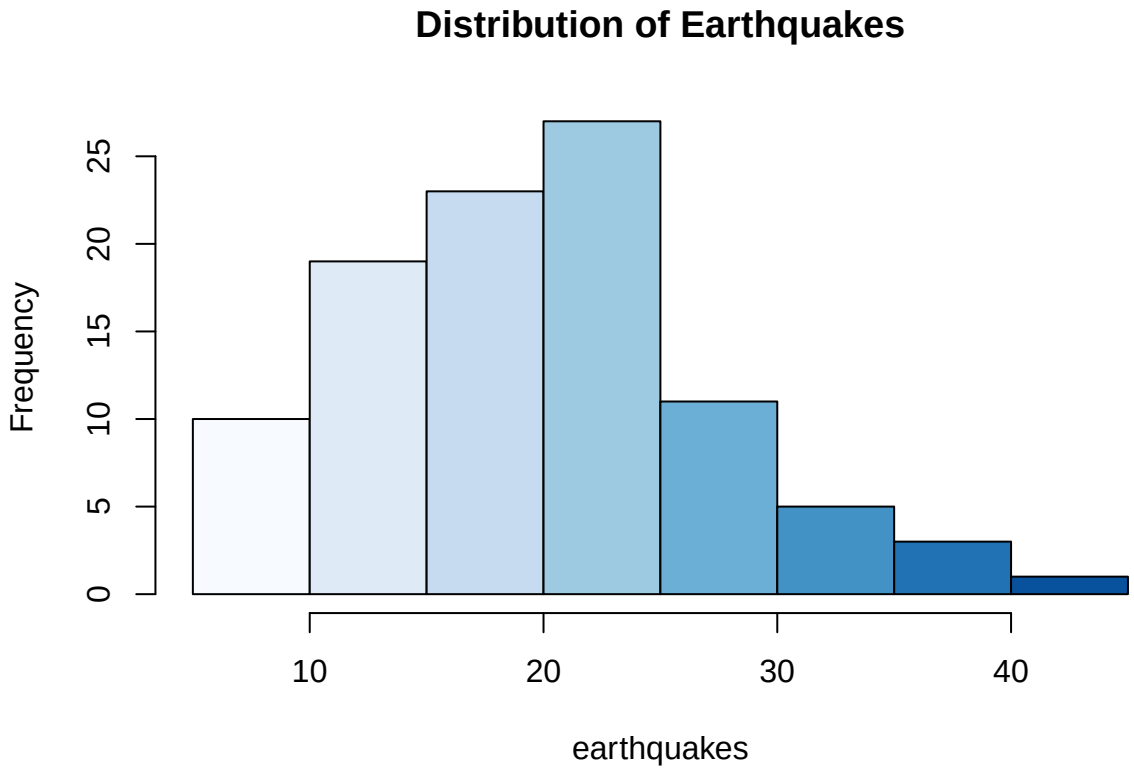
```
##
##  chi-squared test for outlier
```

```
##
## data:  sample
## X-squared = 3.076, p-value = 0.07945
## alternative hypothesis: lowest value 10 is an outlier
```

No outliers exist in the lower boundary (left tail) of our data, as proven by the p-values greatly exceeding 0.05. The case here shows that higher frequencies of earthquakes are unusual.

When reflecting upon the data, we can plot a histogram. The amount of earthquakes can be sorted into bins, and the distribution of data within those margins will show where most of the data will be sorted.

```
hist(data$earthquakes,
  xlab = "earthquakes",
  main = "Distribution of Earthquakes",
  breaks = sqrt(nrow(data)),
  col=blues9
) # set number of bins
```



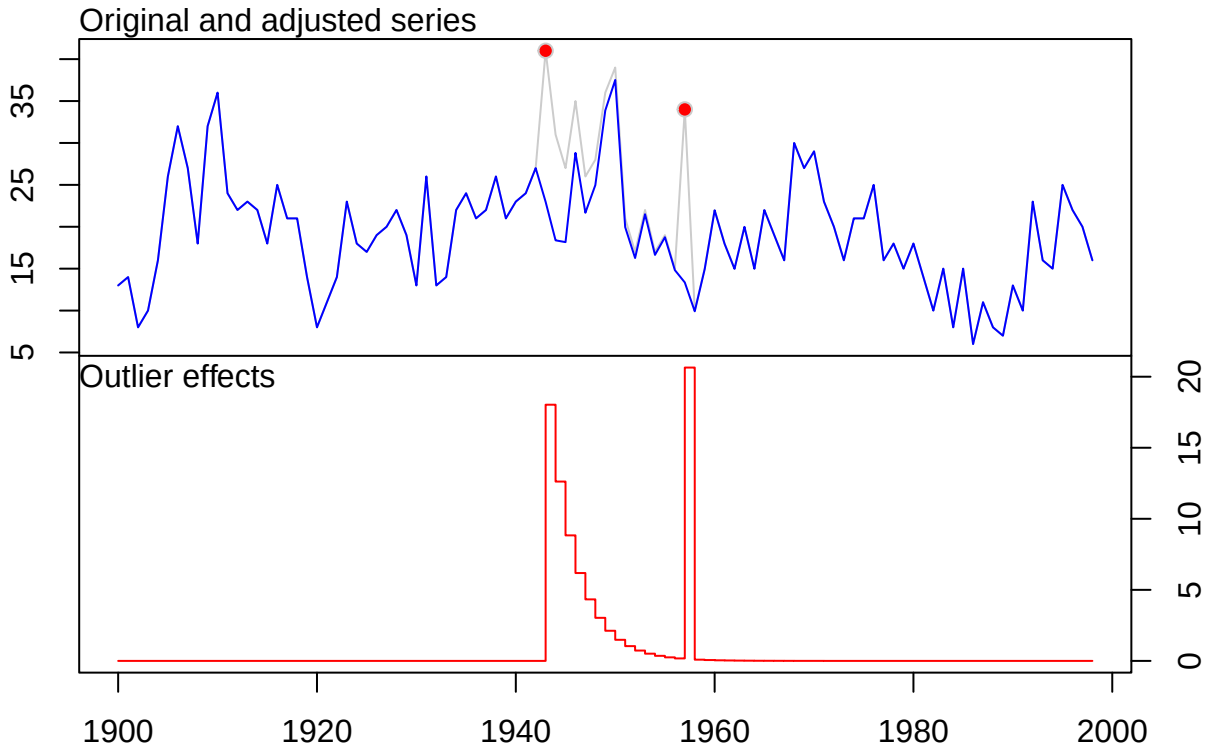
Here, we can see there is a split density of the data. Essentially, most of the data exists between 5 and 25 total earthquakes, with some outliers past 30. Compared to the statistics tests, we see our smaller sample indicates the upper bound of 39 is an outlier. Therefore, we can consider the upper bound of our data having outliers, mainly past 39 being considered outliers. This means any years with more than 39 earthquakes may be considered abnormal.

For time-series related claims as such, we can use `tso()` to test for time-series outliers.

```
tso_table <-tso(ts_data)
tso_table
```

```
## Series: ts_data
## Regression with ARIMA(1,0,0) errors
##
## Coefficients:
##          ar1  intercept     TC44     AO58
##       0.5606    19.0876  18.0287  20.5209
## s.e.  0.0849     1.2045   5.3434   4.6269
##
## sigma^2 = 28.99:  log likelihood = -305.29
## AIC=620.59   AICc=621.23   BIC=633.56
##
## Outliers:
##   type ind time coefhat tstat
## 1   TC  44 1943   18.03 3.374
## 2   AO  58 1957   20.52 4.435
```

```
plot(tso_table)
```



When we viewed our boxplot, we saw most outliers in 1949. One outlier occurred close to 1949 but wasn't registered. This data shows us different types of outliers:

"By default:"AO" additive outliers, "LS" level shifts, and "TC" temporary changes are selected; "IO" innovative outliers and "SLS" seasonal level shifts can also be selected."

Source: TSO Documentation

We can infer that outliers exist in 1943 and 1957, with a t-statistic used to test for them. The visualization highlights these outliers that were initally missing from the boxplot, and are indeed outliers. Furthermore, we can also prove that outliers persist between 1940 and 1960.

This all makes sense, however we may be also concisely detect outliers using a classifier.

Diving Deeper Using an SVM Unsupervised Technique

One-Class SVM Unsupervised Technique

```
model_oneclasssvm <- svm(data,type='one-classification',kernel = "radial",gamma=0.05,nu=0.05)
model_oneclasssvm
```

```
##
## Call:
## svm.default(x = data, type = "one-classification", kernel = "radial",
##     gamma = 0.05, nu = 0.05)
##
##
## Parameters:
##    SVM-Type:  one-classification
##  SVM-Kernel:  radial
##       gamma:  0.05
##          nu:  0.05
##
## Number of Support Vectors:  7
```

Using this technique, we use a SVM to turn linear classifications into a non-linear form, and decisions are created upon those classifications. This helps us to directly identify outliers among data.

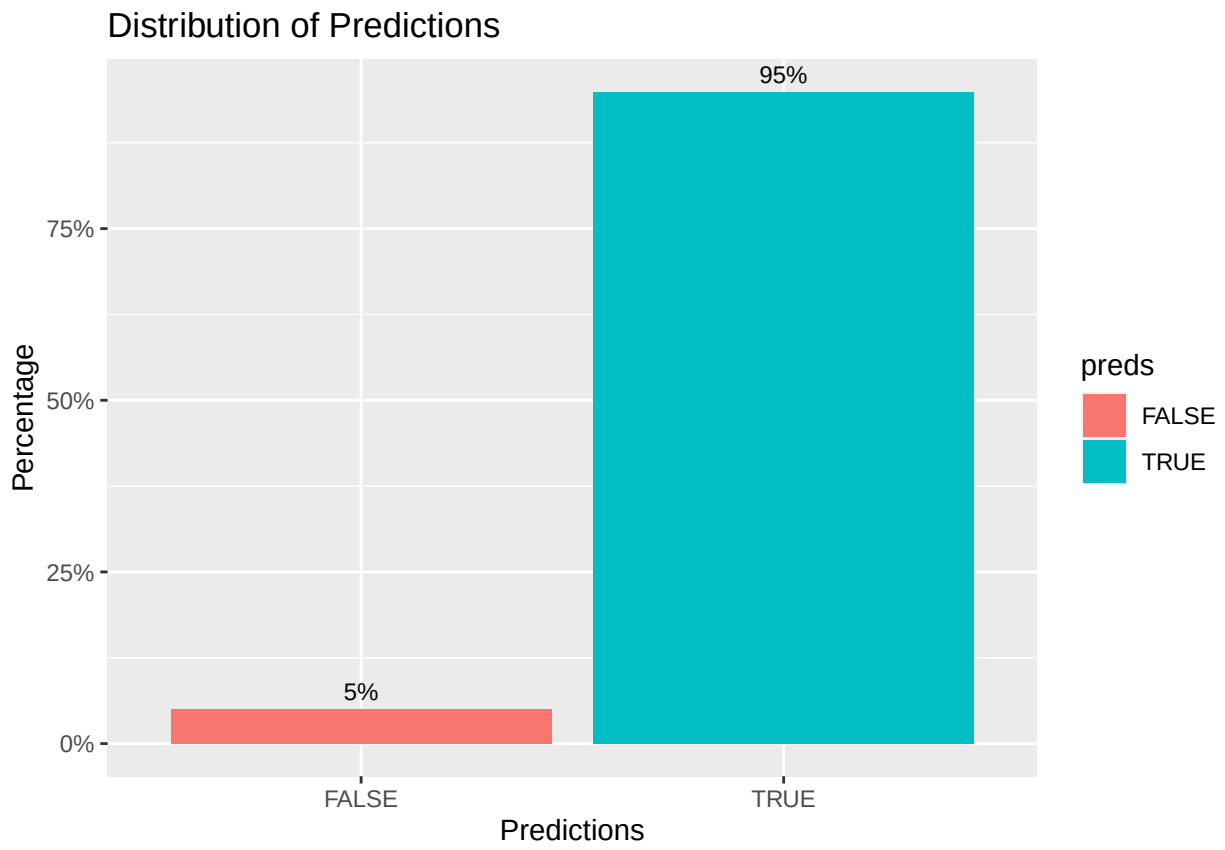
```
preds <- predict(model_oneclasssvm,data)
data$preds <- (preds)
str(data)
```

```
## 'data.frame':    99 obs. of  3 variables:
##  $ year       : int  1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 ...
##  $ earthquakes: int  13 14 8 10 16 26 32 27 18 32 ...
##  $ preds      : logi  TRUE TRUE FALSE FALSE TRUE TRUE ...
```

The values generated will be classified as false if they prove to be outliers. We can count the amount of these values and compare:

```
data %>%
    count(preds = factor(preds)) %>%
    mutate(pct = prop.table(n)) %>%
    ggplot(aes(x = preds, y = pct, fill = preds, label = scales::percent(pct))) +
    geom_col(position = 'dodge') +
    geom_text(position = position_dodge(width = .9),    # move to center of bars
              vjust = -0.5,    # nudge above top of bar
```

```
            size = 3) +
  scale_y_continuous(labels = scales::percent)+
  ggtitle("Distribution of Predictions")+
  xlab("Predictions")+
  ylab("Percentage")
```



There are 5% outliers within the data. We can also look at this in the original data.

```
data %>%
  filter(preds =="FALSE")
```

```
##   year earthquakes preds
## 1 1902           8 FALSE
## 2 1903          10 FALSE
## 3 1943          41 FALSE
## 4 1950          39 FALSE
## 5 1986           6 FALSE
```

The years specifically indicated as outliers were very similar to those located in the boxplot and prior plots, indicating this model has made accurate decisions upon the decision boundary it had created. We notice two outliers at the beginning and end of our data that are a lower value, which indicates there is a normal relationship of these outliers too. That tells us some years lead to heavy earthquake activity, and something may have caused these anomalies worth investigating further.

Conclusion

There are plenty of ways to detect outliers within your data and draw predictions that may display outliers to you. As you consider larger datasets, machine learning classifiers such as an RBF Kernel SVM / Non-Linear SVM become very useful. At face value, we can't always see outliers straight away. With the help of statistics tests, time-series outlier detection (where warranted), and machine-learning algorithms, we can achieve satisfying results.

References

- https://statsandr.com/blog/outliers-detection-in-r/#grubbss-test
- https://www.datavedas.com/anomaly-detection-in-r/
- https://www.datacamp.com/community/tutorials/support-vector-machines-r
- https://datascienceplus.com/outliers-detection-and-intervention-analysis/