IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Anomaly detection methods and tools for big data[1]

Shir Kamenetsky Yadan,
Bank of Israel

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Anomaly Detection Methods and Tools for Big Data

Shir Kamenetsky Yadan,
Bank of Israel

October 2021

Abstract

Anomaly detection is the process of identifying observations in a dataset which deviate from the norm. In central banking, anomaly detection plays an essential role in managing, monitoring, and analysing data repositories. While traditional anomaly detection is manual, detecting anomalies in big data is humanely impossible. Consequently, the development of tools for mechanized and efficient detection of anomalous observations is becoming increasingly crucial for the ongoing work of database managers and researchers as real time big data repositories continue to expand at an accelerated rate.

This paper presents a customized RShiny dashboard built using R's Flexdashboard format in Rmarkdown for user defined anomaly detection. The dashboard uses the Forex Data Repository which consists of daily transactions in foreign exchange derivatives and interest rates executed in the OTC market by financial intermediaries in Israel and abroad. These daily transactions accumulate to millions of records a year across 40 variables. In order to tackle the challenge of conducting quality control as well as analysing and extracting useful insights from a database of this size, we developed a tool for detecting anomalies which includes three main features; 1) Data upload and pre-processing, 2) Traditional anomaly detection, 3) Univariate and multivariate non-parametric anomaly detection .In the paper we expand on each one of these features and demonstrate their application. Some of the traditional methods we include are a visual examination of the data distribution and implementation of variance stabilizing and normalizing techniques such as Box-Cox transformation, subsequently applying standard deviation, median absolute deviation and interquartile range for detecting outliers. Economic series such as Forex transaction amounts are often characterized by highly right-skewed distributions making it difficult to use such traditional techniques.

Applying transformations to the data is not always enough to meet the symmetry or other parametric assumptions required by many of the traditional methods. In such circumstances a distribution-free test for outliers in data drawn from an unknown data generating process may give more reliable results. We apply two innovative non-parametric methods integrated into the third feature of the tool; a bootstrapping procedure for outlier detection Bootlier Plot (Singh and Xie, 2003) and Isolation Forests (Liu, Ting, and Zhou, 2009), and show their implementation on Forex data. Finally, we discuss the potential use of similar anomaly detection tools in different big data repositories such as the Central Credit Register and Payment Systems repository.

Contents

1. Data Upload, Pre-processing, and Exploration

Throughout the following sections we will analyse a dataset of about 290,000 Forex transactions from the second half of June 2020. The first step after uploading the dataset is choosing a variable to analyse and choosing desired filters as seen in Figure 1. Initial exploratory analysis as seen in Figures 2, 3, and 4 is important before proceeding to anomaly detection. This window of the tool offers tabular and visual exploration of the data which allows for familiarization with data distribution characteristics. This is necessary for choosing an appropriate anomaly detection method.

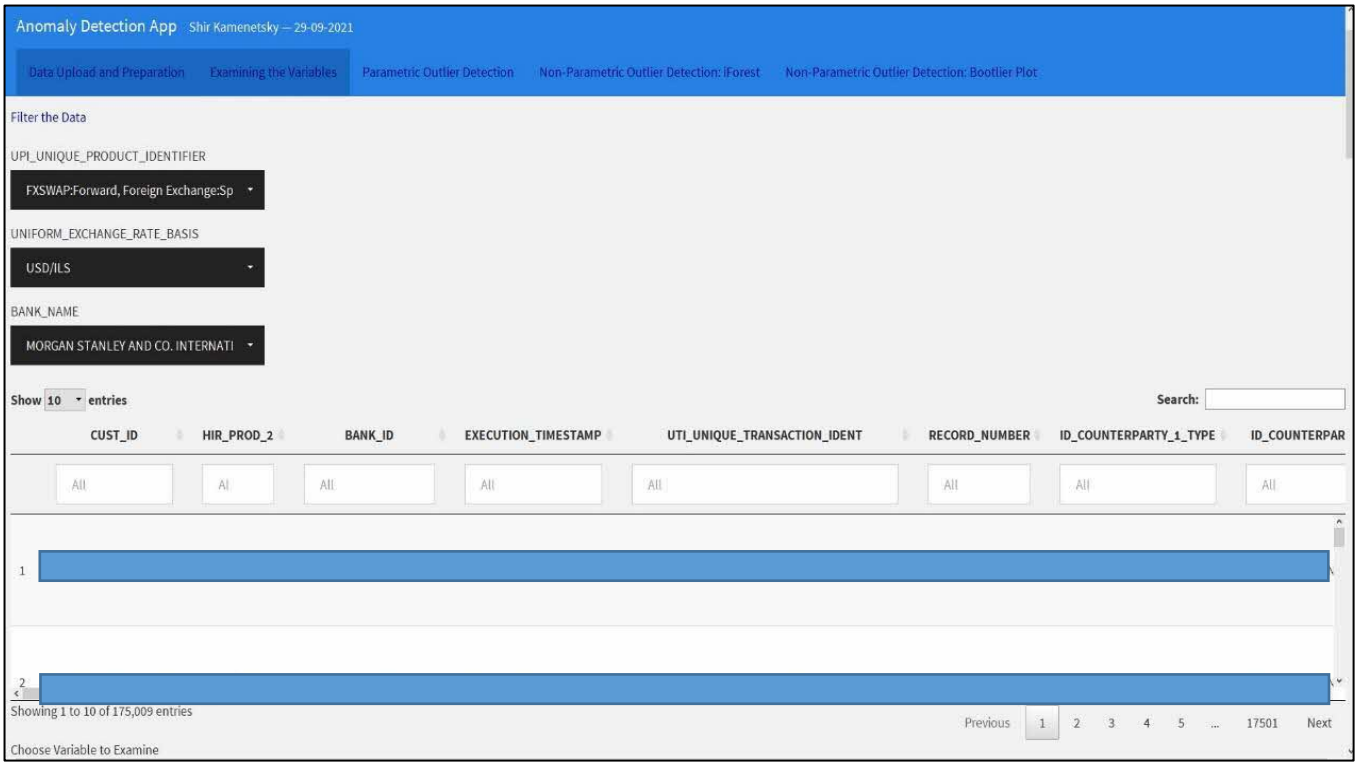


Figure 1: Filtering Window

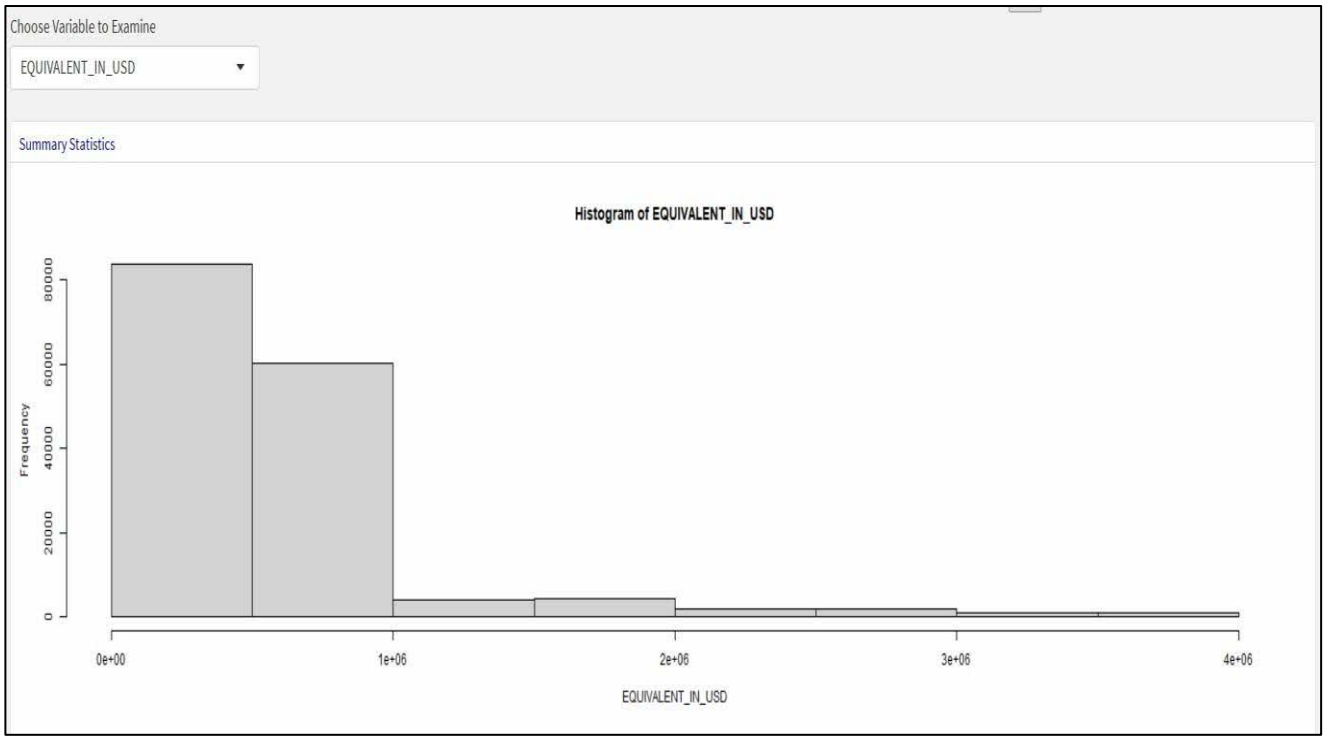


Figure 2: Histogram

In this example we choose to examine the variable "EQUIVALENT_IN_USD" which is the foreign exchange transaction amount in US dollars. We filter only the exchange basis, choosing transactions between US dollars and Israeli Shekel. We choose all UPI's (unique product identifiers) and all banks leaving us with about 175,000 entries. On the choice of variable window shown in Figure 2, a histogram appears giving us an idea of the distribution characteristics. Like in many financial data, the distribution has a strong right skew. This observation is important for further analysis. Figures 3 and 4 show additional visualizations of the data including a scatterplot of the chosen variable over time, with the option to choose a categorical variable for the colour of the points making the plot 3-dimentional. In this example each point is a single transaction sum with transaction time on the x-axis and colour by UPI. This plot can point out potential anomalies in the context of time. An additional plot called a ridge plot is shown in Figure 4, once again with the option of choosing a categorical variable to filter by. This time categories are separated on the y-axis while the x-axis shows the numerical variable of interest, in our case "EQUIVALENT _IN USD". What we learn from this plot is the

distribution shape of "EQUIVALENT IN USD" for each of the different UPI's making it visually easy to compare between distributions of different UPI's. We notice, for example, that some UPI's have unimodal distributions while others are multi-modal.

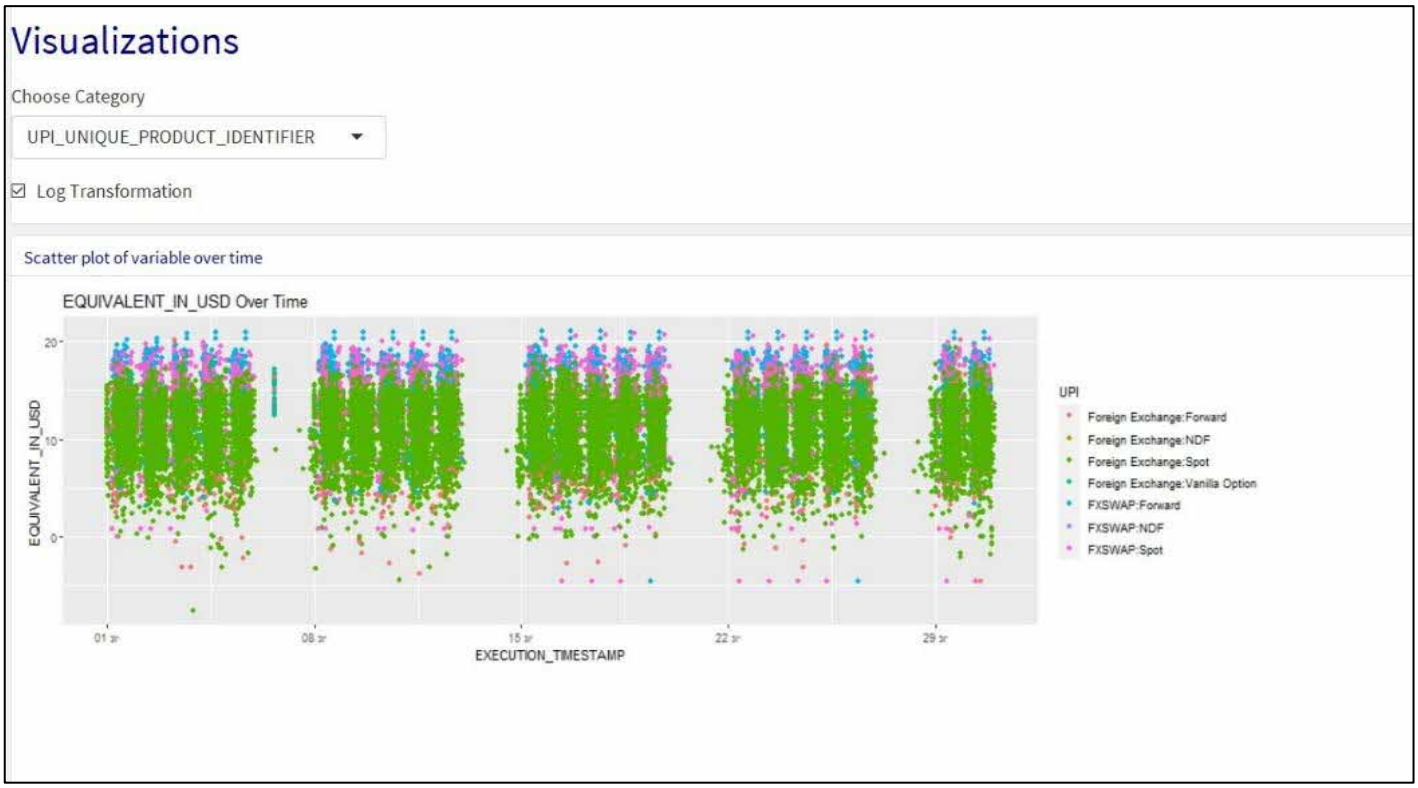


Figure 3: Scatterplot

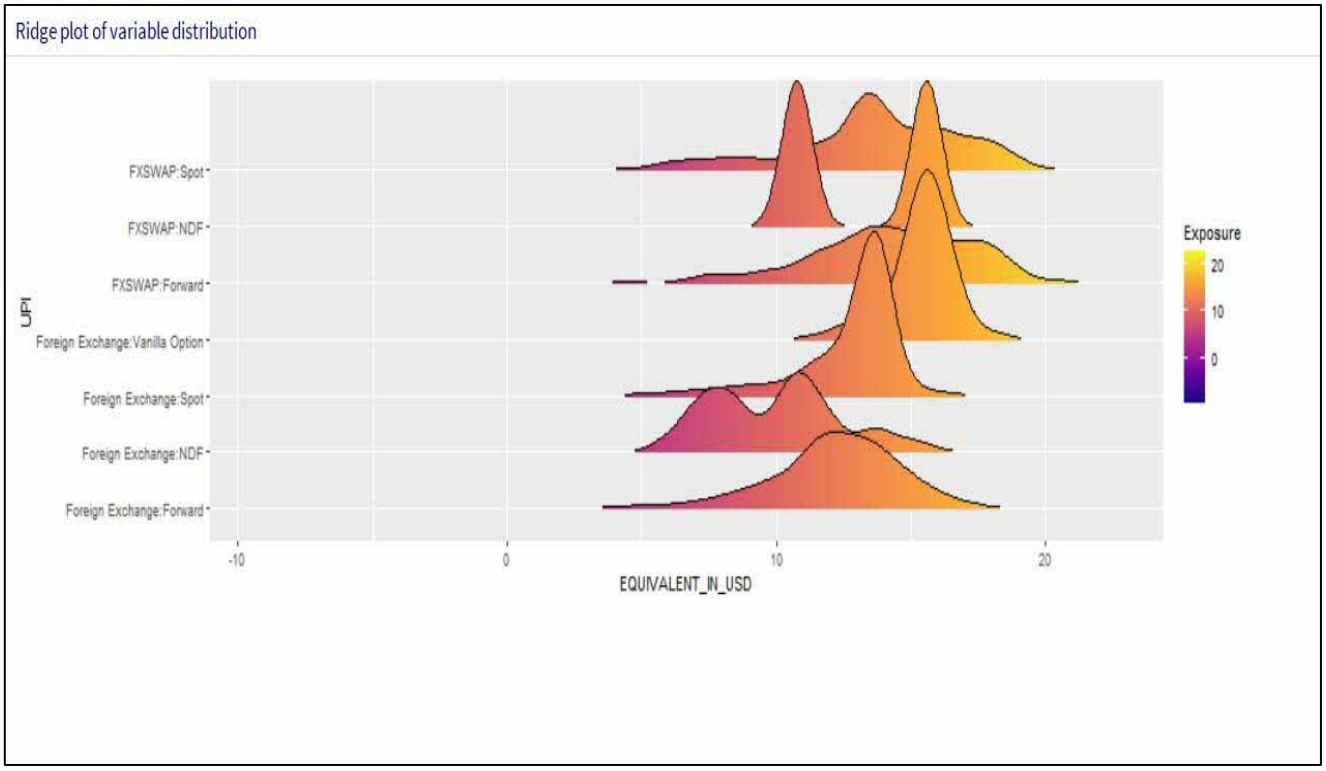


Figure 4: Ridge Plot

2. Traditional Methods

Traditional methods for identifying anomalies such as setting thresholds by taking a number of standard deviations from the mean are often most reliable when data meets certain parametric assumptions such as symmetry or normality of the distribution. We have seen in the previous section that our data, much like any financial data, is neither normal nor symmetric. In order to use traditional methods for anomaly detection on such data we begin by applying transformations to attempt to bring the data distribution to a more symmetric and normal shape. Two transformations are offered in this tool; Natural Log transformation and Box-Cox transformation.

Box-Cox Transformation

Box-Cox attempts to approximate the normal distribution.

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} \text{ if } \lambda \neq 0 \\ \ln y_i \text{ if } \lambda = 0 \end{cases}$$

Where $y_i^{(\lambda)}$ is the Box-Cox transformed data and the optimal λ is one which results in best approximation of normal distribution curve.

After transforming the data, threshold values for anomalous data are calculated using 4 different choices of center and variability metrics; 1) Mean and Standard Deviation, 2) Median and Median Absolute Deviation (MAD), 3) Median and Double MAD, 4) Inter-Quartile Range.

Mean and Standard Deviation

Mean and Standard deviation is common practice and the most parametric as well as non-robust method. It is most reliable with normally distributed data. In the case of normally distributed data, 3 standard deviations taken from both sides of the mean will cover 99% of the data and any observations outside of these thresholds can be considered anomalies.

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

$$Anomaly\ Threshold = Mean \pm \alpha * SD$$

Median and MAD

The Median and MAD method is a non-parametric method and robust in two main aspects; the first being the use of the median as a measure of centrality which is robust to outliers in itself, and once again the use of the

median for aggregation in the MAD measure of variability. Additionally, unlike standard deviation which squares the deviations of each observation from the mean causing larger deviations to explode, MAD uses absolute value which minimizes the effect of large deviations and thus adds to robustness.

$$MAD = k * median(|Y_i - median(Y)|)$$

$$Anomaly\ Threshold = Median \pm \alpha * MAD$$

Where k is called the scale factor and is taken to be 1.4826 if data is normally distributed. In this case MAD can be used as a consistent estimator for the estimation of the standard deviation. Since the distribution of our data is unknown we take k to be 1.

Median and Double MAD

The classic Median and MAD method defines a symmetric interval of anomaly thresholds around the median. This works best when the distribution is indeed symmetric. In cases like ours where the distribution is heavily skewed, often even after applying a transformation, there is the Double MAD measure which calculates two separate MAD values for the left and right sides of the distribution (using median as the center). [3]

$$\tilde{Y} = median(Y)$$

$$Y^{(u)} = \left\{ y | y \in Y \bigcap y \geq \tilde{Y} \right\}$$

$$MAD^{(u)}(Y) = k * median(\left| Y_i^{(u)} - \tilde{Y} \right|)$$

Where one again k is the scale factor. u indicates "upper" distribution observations -those which are greater or equal to the median. The same calculations are done using lower observations for $MAD^{(l)}(Y)$. We now take α upper deviations and α lower deviations from the median to get upper and lower anomaly thresholds.

$$Lower\ Threshold = median - \alpha * MAD^{(l)}(Y)$$

$$Upper\ Threshold = median + \alpha * MAD^{(u)}(Y)$$

IQR and Tukey's Fences

Lastly, the anomaly detection tool gives the option of using IQR as a measure of variability with Tukeys' Fences to calculate anomaly thresholds. This method is nonparametric and robust.

$$IQR = Q_3 - Q_1$$

Where Q3 is the value that holds 25% of the values above it and Q1 is the value that holds 25% of the values below it.

$$Lower\ Threshold = Q_1 - \alpha * IQR$$

$$Upper\ Threshold = Q_3 + \alpha * IQR$$

In this case, α is taken to be 1.5 and the choice of alpha in the anomaly detection tool is disabled.

Figure 5 shows a comparison of all four methods on our example dataset. Notice that in this window of the anomaly detection tool the user chooses a transformation method, measure of centrality and variability pair, number of deviations to take from the center (alpha), and whether they want to consider upper, lower, or all outliers. A histogram is then plotted with dashed vertical lines showing the center value and anomaly thresholds and a kernel density curve of the normal distribution using either the mean and standard deviation or the median and MAD to generate observations from the normal distribution. In this example we choose alpha to be 4. We see how standard deviation thresholds are placed much further from the center than thresholds of the other measures. This occurs due to the lack of robustness of the standard deviation. The variability is affected by the values at the tails of the distribution more so than the robust measures. Notice the symmetric thresholds of the standard deviation and MAD methods versus the non-symmetric thresholds of Double MAD and IQR which are pulled further away from the center on the left side of the distribution due to the longer left tail. Choosing the method and size of alpha are ultimately up to the user. It is important that the user has an expertise in the field and good familiarity and knowledge of the data. The user can then decide which method gives the most accurate results and whether the flagged observations are indeed anomalous points.

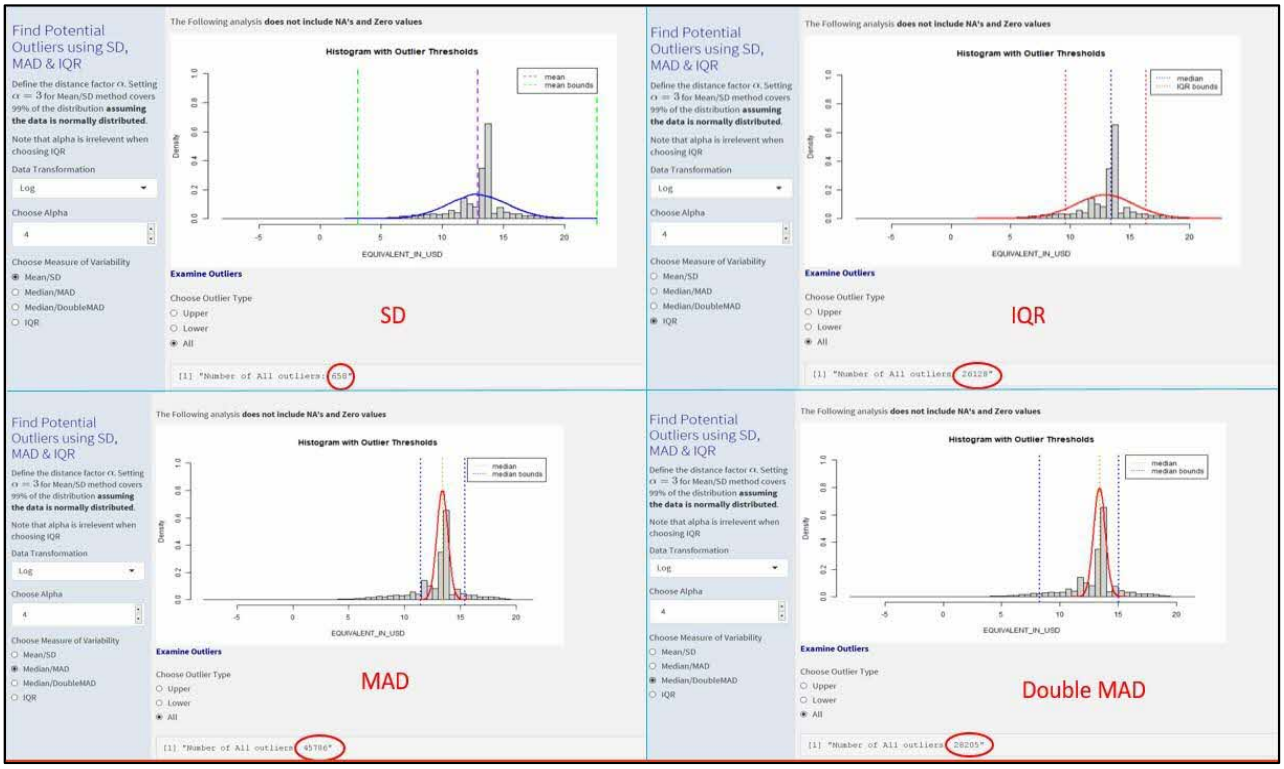


Figure 5: Anomaly Detection

3. Nonparametric Univariate and Multi-Variate Methods

We may prefer to use nonparametric methods on data drawn from an unknown data generating process rather than traditional methods on transformed data. We offer two distribution-free methods for anomaly detection; 1) Isolation Forest and 2) Bootlier Plot. Both methods have multivariate implementation; however, we do not currently offer multivariate Bootlier Plot in our anomaly detection tool. In this section we will explain the methods and demonstrate their use in the anomaly detection tool.

3.1. Isolation Forest (Liu, Ting, and Zhou, 2009)

The basic idea of this method is to isolate anomalies rather than profiling normal instances. [2] Anomalies are "few" and "different" making them more susceptible to isolation than normal points. Isolation is conducted via construction of tree structure. Partitions are generated by randomly selecting a variable and then randomly selecting a split value between the maximum and minimum values of selected variable. This is done iteratively until each observation is isolated to its own node. Path length is equivalent to the number of partitions required to isolate a point, or in tree structure the number of edges from root node to external terminating node. Observations that are quicker to isolate and have shorter path lengths are ones considered to be anomalies. We see an example of this in Figure 6 where x_0 is an obvious anomaly placed far from the other points and is therefore isolated with fewer partitions.

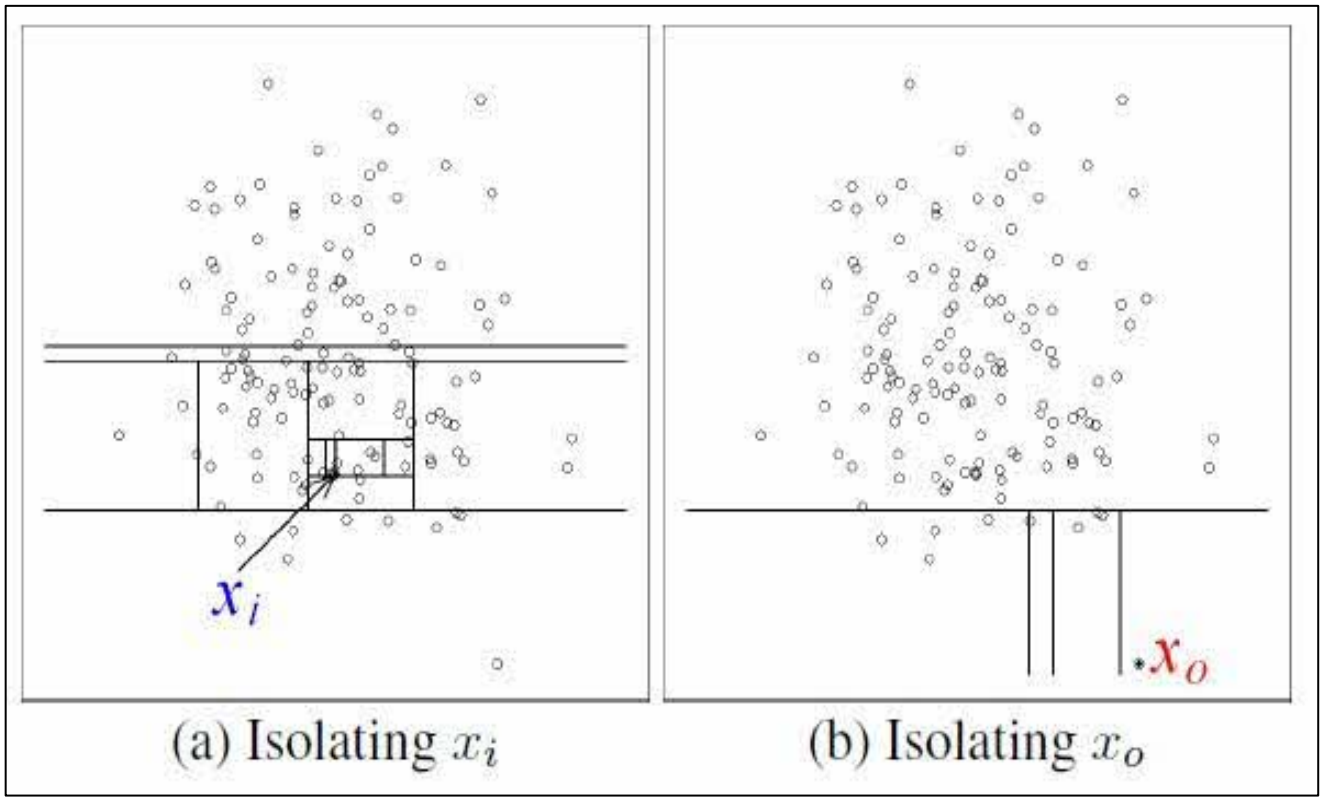


Figure 6: Isolation Forest

These steps are applied to subsets of the data on an ensemble of trees called an Isolation Forest. Path lengths for each observation are aggregated over the trees and a final anomaly score is given to each observation. This score is between 0 and 1 where scores close to 1 are considered anomalies and scores smaller than 0.5 can be considered regular instances. In figure 7 we see a visual example of an isolation forest with anomalous points closer to the root of the tree.

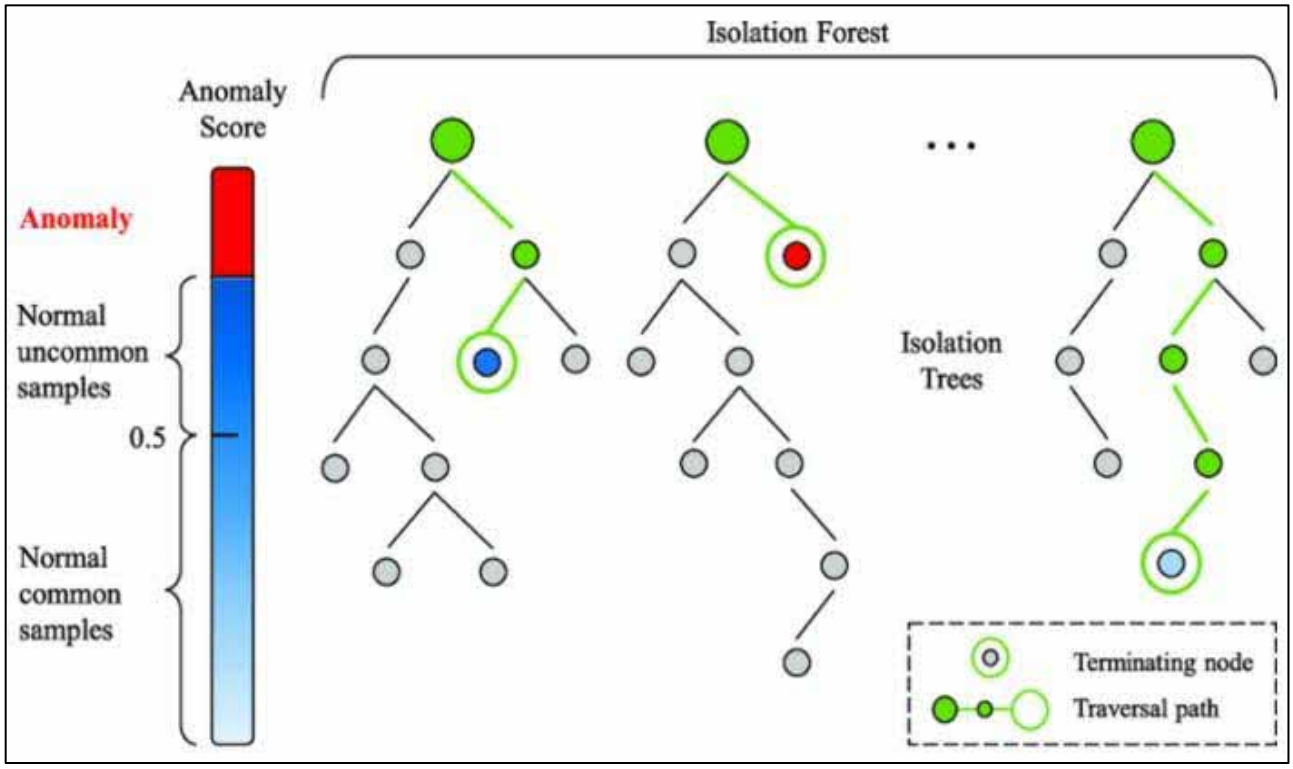


Figure 7: Isolation Forest

A major benefit of Iforest over other unsupervised anomaly detection methods resides in the use of sub-sampling with relatively small samples for each tree. Sub-sampling is conducted by random selection of instances without replacement. Building Isolation trees on smaller samples of data reduces the swamping and masking effects which are common in other anomaly detection

methods[1]. Swamping occurs when normal observations are too close to anomalous points and are wrongly classified as anomalies. In other words the normal points are ”swamped” by the anomalies. In masking, a group of anomalous observations close together ”mask” their own presence and they are classified as normal points. These effects are common especially in very large datasets. For this reason Iforest is an especially suitable method for anomaly detection in our large data repositories.

Figure 8 demonstrates the use of Iforest in our anomaly detection tool on our example data. The user may choose multiple variables for analysis, in this example transaction amount, UPI, exchange rate, and coin are chosen. The user then chooses an anomaly threshold - the anomaly score for which any value above this score will be considered an anomaly. In this example an observation is classified as an anomaly if it receives an anomaly score over 0.75. The user can then run the algorithm and view a table of all points and their classification as anomalous or normal, a quantile table of the anomaly scores, and the final number of anomalies in the data.

[1] For extended explanation of how Iforest handles swamping and masking see Liu, Ting, and Zhou 2008

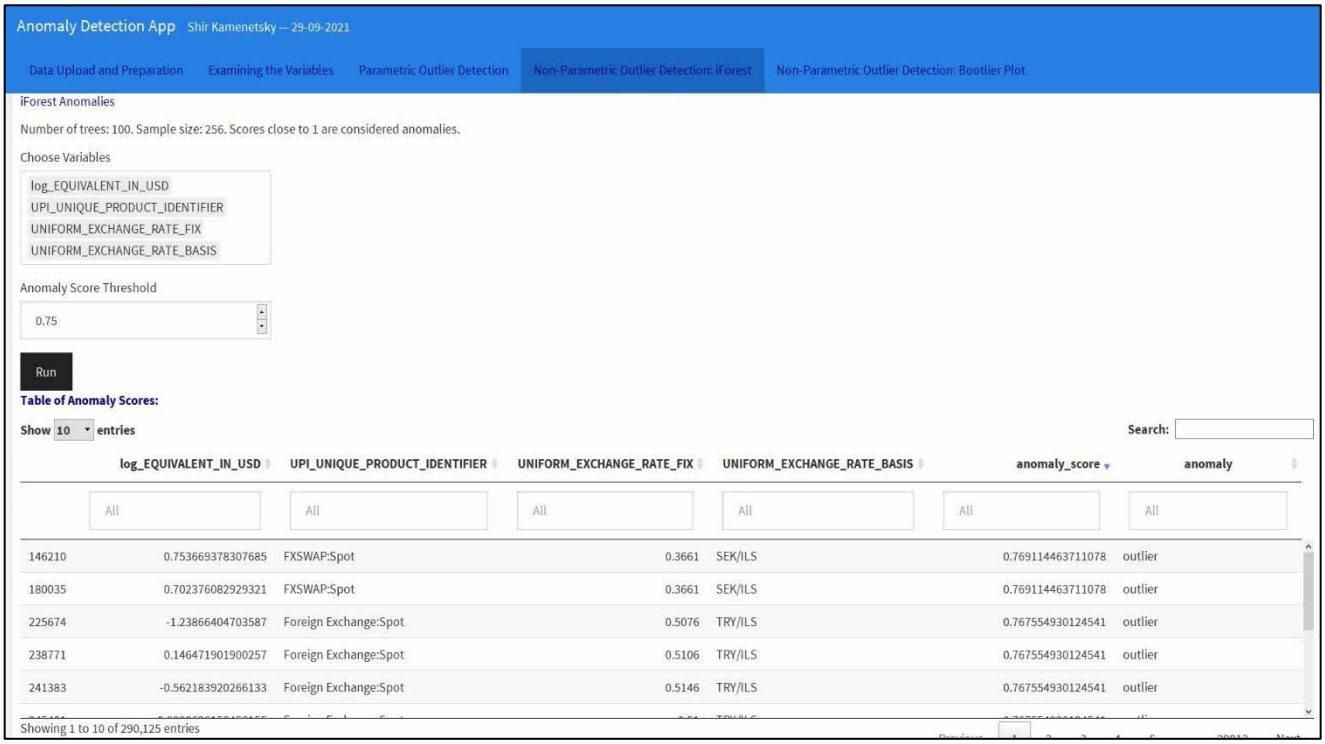


Figure 8: Isolation Forest

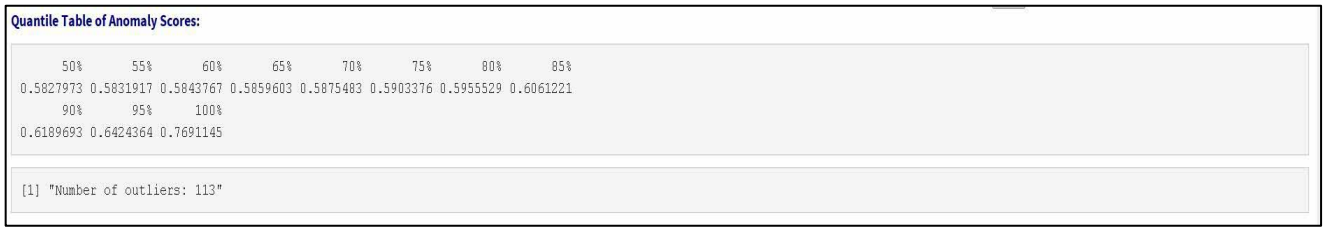


Figure 9: Isolation Forest - Quantile Table

3.2. Bootlier Plot (Singh and Xie, 2003)

The Bootler Plot method [4] is based on bootstrapping. When an outlier exists in a dataset, some bootstrap samples will contain the outlier while others will not. The presence of an outlier is expected to cause a significant increase or decrease in the bootstrap mean, and make the bootstrap distribution of the sample mean a mixture distribution. Therefore, we expect the histogram of the

sample mean to be multimodal. In order to make the bootstrap histogram more sensitive to a potential outlier, the chosen bootstrap statistic is the "mean – trimmed mean" [2]. Where the trimmed mean is the mean of the bootstrap sample after trimming k observations from each side of the sorted sample.

Mean-Trimmed Mean Statistic:

$$T(Y^*) = \frac{1}{n}\sum_{1}^{n} Y_i^* - \frac{1}{n-2k}\sum_{k+1}^{n-k} Y_{(i)}^*$$

Where $T(Y^*)$ is the mean-trimmed mean statistic, $Y_1^*, Y_2^*, \ldots, Y_n^*$ denote bootstrap draws from a certain bootstrap, and Y_i^* the corresponding order statistics.

In Figures 10 and 11 we see the Bootlier Plots of data, and the same data with an additional anomalous observation. Figure 11, with the anomalous point has an obvious additional "bump". Bootlier plot becomes less practical when data is large and many potential

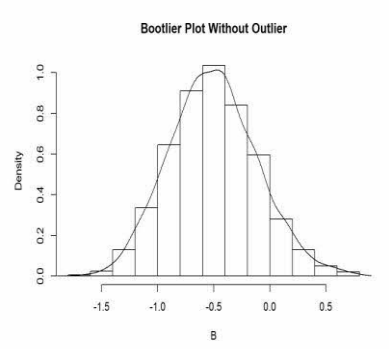


Figure 10:

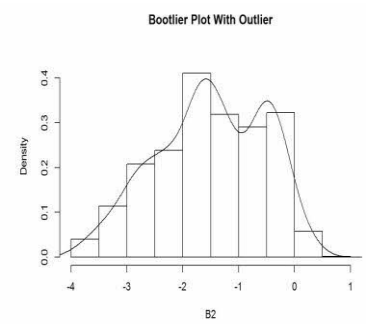


Figure 11:

outliers exist. Singh and Xie refer to this issue in their paper. The probability of having a bootstrap sample of a large size, *n*, free from potential outliers is very small and we may not see a clear bump in Bootlier plot. The solution is to reduce the bootstrap sample size to a fraction of *n*, i.e., [αn], α∈(0,1]. A practical

[2] The reasoning for greater sensitivity of "mean-trimmed mean" statistic to potential outliers is explained in further detail in Singh and Xie, 2003

recommendation is to look at several Bootlier plots at different bootstrap sample sizes – if any one is found bumpy that would be indicative of the presence of outliers. We find large data size to be an issue when using Bootlier Plot on our data. Even when following the recommendations of the article for large amounts of data we typically get unimodal histograms. Since we do not have a set of "real" anomalies in our data to compare with, it is difficult to know whether this outcome stems from the fact that there really are no anomalies in the data or is a technical outcome of the algorithm such as the one discussed regarding very large data sets. In any case we offer the user the flexibility of choosing the number of desired bootstraps, sample size, and trimming amount[3] and view Bootlier Plots for each parameter trio. Another problem with large datasets in this method is a slower running time. It may be time-impractical to try very many parameter combinations especially when choosing large numbers of bootstraps.

The Bootlier Plot indicates whether a dataset holds anomalies or not, however; it does not identify and give the values of these outliers. Candelon and Metiu, 2013 [1] extend on the Bootlier Plot in the Deutsche Bundesbank Discussion Paper – "A distribution-free test for outliers" developing a method for identifying outliers from the Bootlier Plot. They term this method the "Bootlier Test". The method uses a two-step process:

1. Test for multimodality: **H0** – Bootlier plot has precisely one mode (and no local minimum), **H1** – Bootlier plot has more than one mode. Test hypothesis using "Bootlier Test" - Bootlier plot coupled with distribution free test for multimodality proposed by Silverman (1981)

2. Identify Outliers: 1) Build subsamples by sequentially cancelling observations from the tails of the original sample ordered in ascending order. 2) Perform Bootlier test on each ordered subsample until the null hypothesis of unimodality cannot be rejected for a particular subset of observations. 3) Data points not contained in this subset are the anomalies.

In figures 12 and 13 we demonstrate the use of Bootlier Plot and Bootlier Test on our example data. In the first example we choose 1000 bootstraps of sample size 2000 (recall that our data size is 175,000 so this is about 1% of the data) and trim amount of 10. With these parameters the Bootlier Plot is

[3] The writers mention that the optimal choice of trim amount is not given theoretical groundwork in the paper, we therefore leave this open for the users to try different sizes

Figure 12: Bootlier Plot

unimodal and there is no evidence of anomalies. Likewise, no anomalies are found using the Bootlier Test. In the second example, we change thee sample size to 100 (about .06% of the data) and trim amount to 2. Now we see a multimodal Bootlier Plot along with one anomaly found by the Bootlier Test. In comparison to the Iforest method as well as the traditional methods we implemented it is evident that the Bootlier Plot and Test find dramatically less anomalies in the data. We must "force" anomalies out of the method by using extreme parameters. This could indicate that we do not have anomalous points in our data, or be caused by a technicality of the method - perhaps because of our data's large size as we mentioned before. For this reason, as well as greater processing power required for running the Bootlier Plot especially for larger bootstrap sizes, we find this method to be less practical and less reliable for our data than the other methods offered in the anomaly detection tool.

Figure 13: Bootlier Plot

4. Concluding Remarks

The anomaly detection tool for big data enhances the efficiency of the ongoing work of database managers. It gives database managers the ability to filter large amounts of data, study data characteristics and distributions via tables and graphs, and signal suspicious observations with the flexibility of choosing between multiple anomaly detection methods, rather than scroll through enormous excel spreadsheets and eyeball data values. It is important to keep in mind and to emphasize to database managers that these methods alone cannot tell them whether an observation is an anomaly or not, but rather point their attention to suspicious observations. The expertise of the database manager is in the field and experience and knowledge with the data is crucial

in deciding whether an observation is indeed anomalous, or not. This tool is made relatively generic and can be implemented on other data repositories with few adjustments. Besides the Forex repository, we are currently working on adapting the anomaly detection tool to the Central Credit Register and the Payment Systems repository.

References

[1] Bertrand Candelon and Norbert Metiu. "A distribution-free test for outliers". In: (2013). url: http://hdl.handle.net/10419/68604.

[2] Fei Tony Liu, Kai Ting, and Zhi-Hua Zhou. "Isolation Forest". In: (Jan. 2009), pp. 413–422. doi: 10.1109/ICDM.2008.17.

[3] Peter Rosenmai. *Using the Median Absolute Deviation to Find Outliers*. url: https://eurekastatistics.com/using-the-median-absolutedeviation-to-find-outliers/. (accessed: 10.11.2021).

[4] Kesar Singh and Minge Xie. "Bootlier-Plot: Bootstrap Based Outlier Detection Plot". In: *Sankhy¯a: The Indian Journal of Statistics (2003-2007)* 65 (2003), pp. 532–559. doi: 10.2307/25053287.

Anomaly Detection Methods and Tools for Big Data

SHIR KAMENETSKY

BANK OF ISRAEL

OCTOBER 22, 2021

Road Map

1. Motivation
2. Data exploration: pre-anomaly detection
3. Traditional Methods
4. Non-Parametric Univariate and Multivariate Methods

Motivation

Role of anomaly detection in central banking:

Role 1

- Managing and monitoring data repositories
  - Quality assurance – find erroneous data observations
  - Alert for sudden changes in economy, deviations from trends

Role 2

- Analyzing data repositories
  - Gaining greater familiarity and deeper understanding of data

Motivation

PROBLEM

Real time **big data** repositories continue to expand at an **accelerated rate**.

Traditional **manual** anomaly detection is **humanely impossible**.

Motivation

SOLUTION

Development of **mechanized** and **efficient** tools for anomaly detection.

No more spreadsheets!

Motivation

Tool: Anomaly Detection Dashboard App

Data: Forex data repository:

- Daily transactions in foreign exchange derivatives and interest rates executed in OTC market by financial intermediaries in Israel and abroad.
- Millions of records a year across 40 variables

Technologies: Shiny from RStudio, flexdashboard, R Markdown

Motivation

Step 1:
Upload data

Step 2:
Choose variables to analyze and add filters

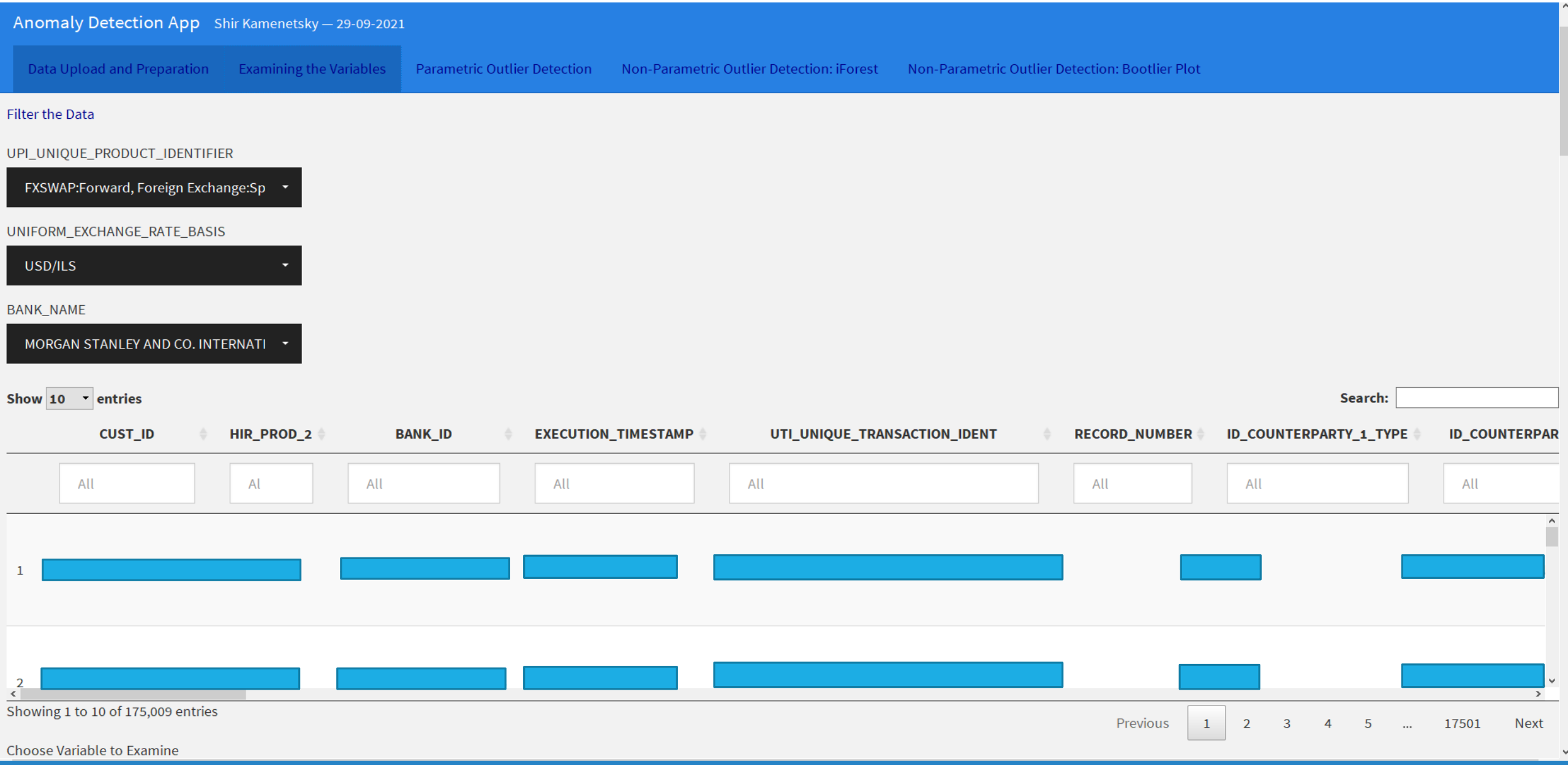
Step 3:
Tabular and graphical initial exploration of data

Tool Flow

Step 4:
Anomaly detection –parametric, non-parametric, univariate, multivariate methods

Data exploration: pre-anomaly detection

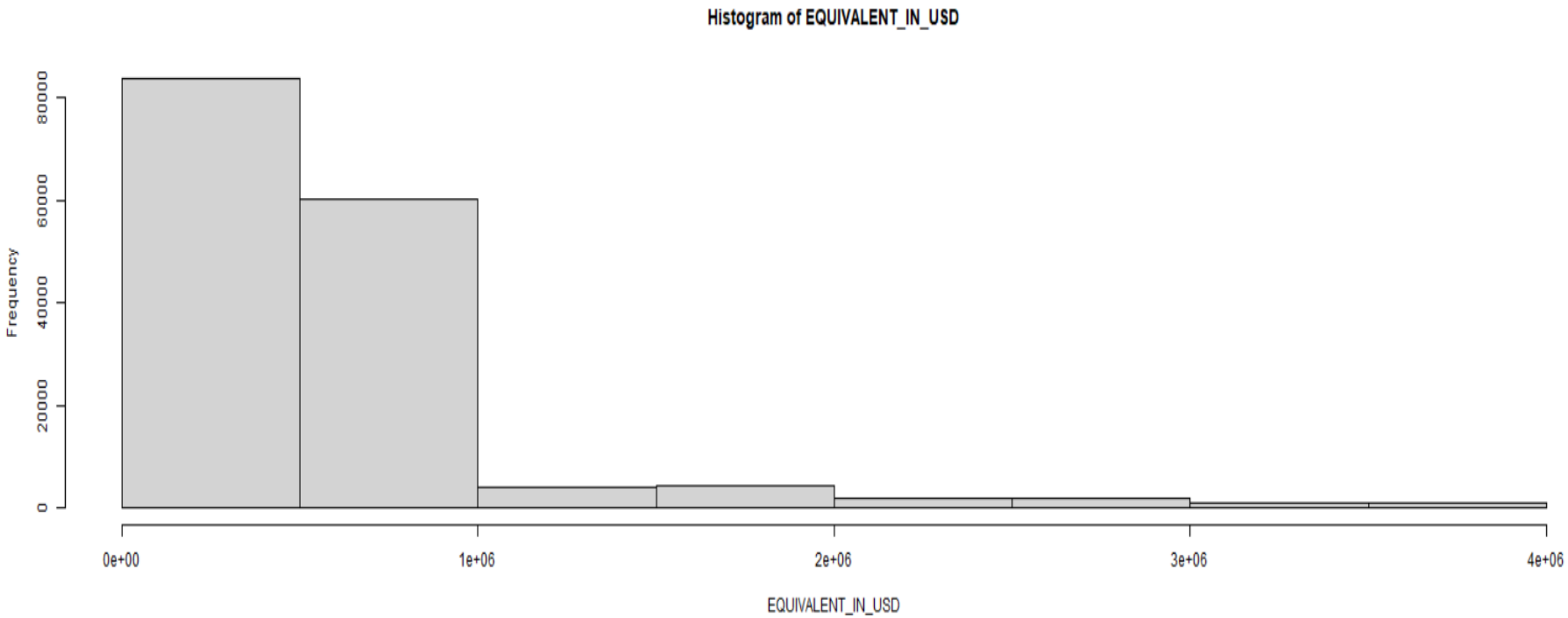
Data exploration: pre-anomaly detection



Choose Variable to Examine

EQUIVALENT_IN_USD

Summary Statistics



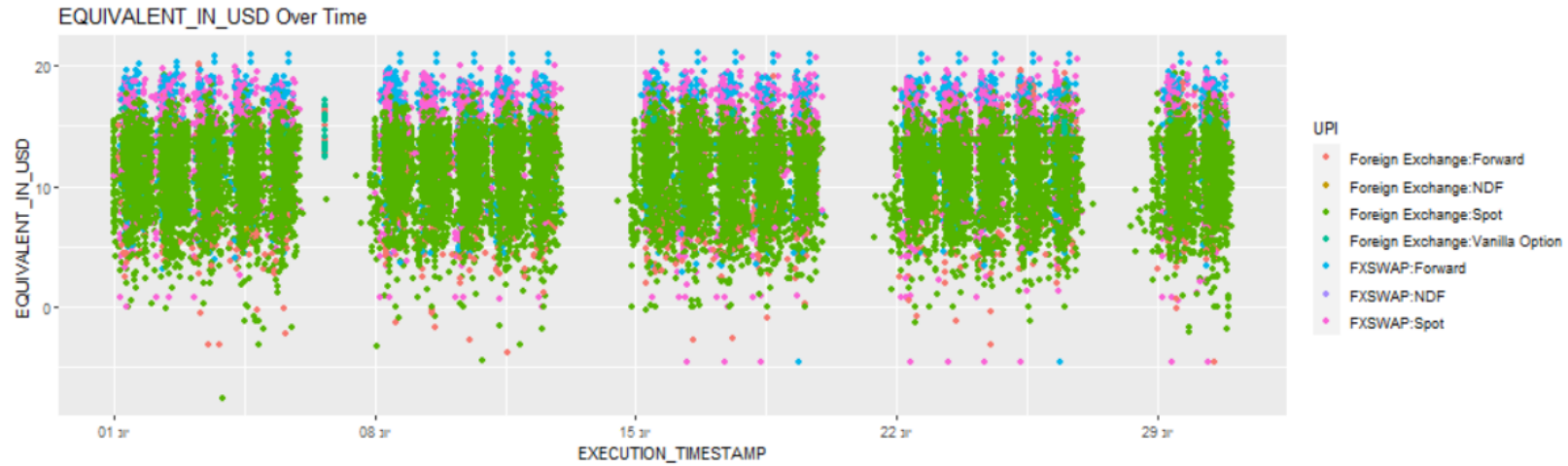
Visualizations

Choose Category

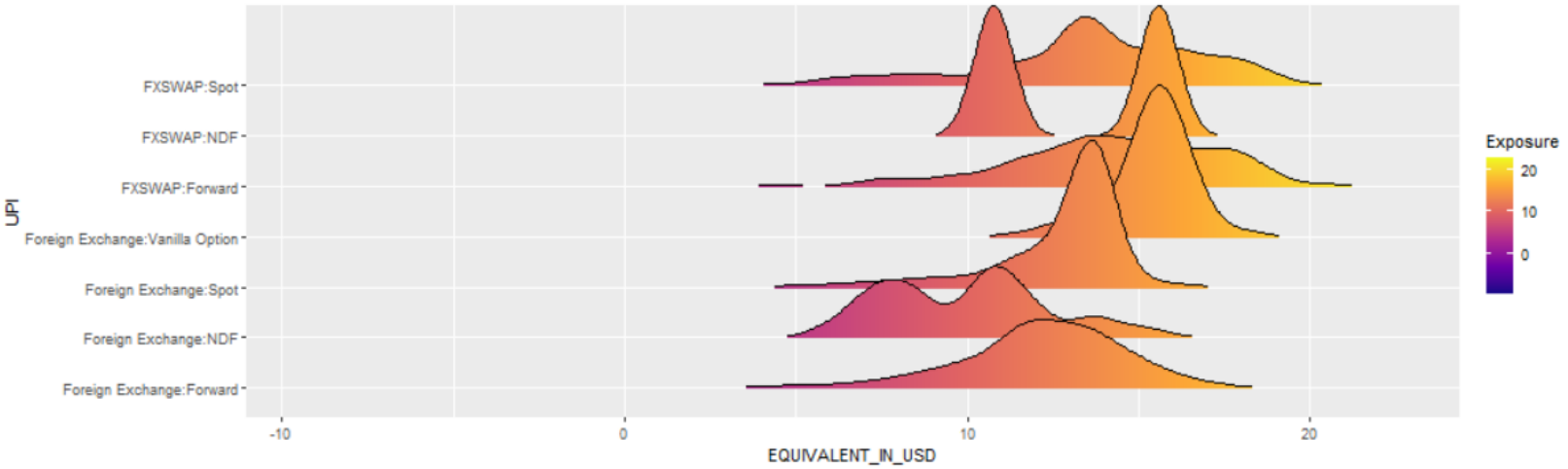
UPI_UNIQUE_PRODUCT_IDENTIFIER

☑ Log Transformation

Scatter plot of variable over time



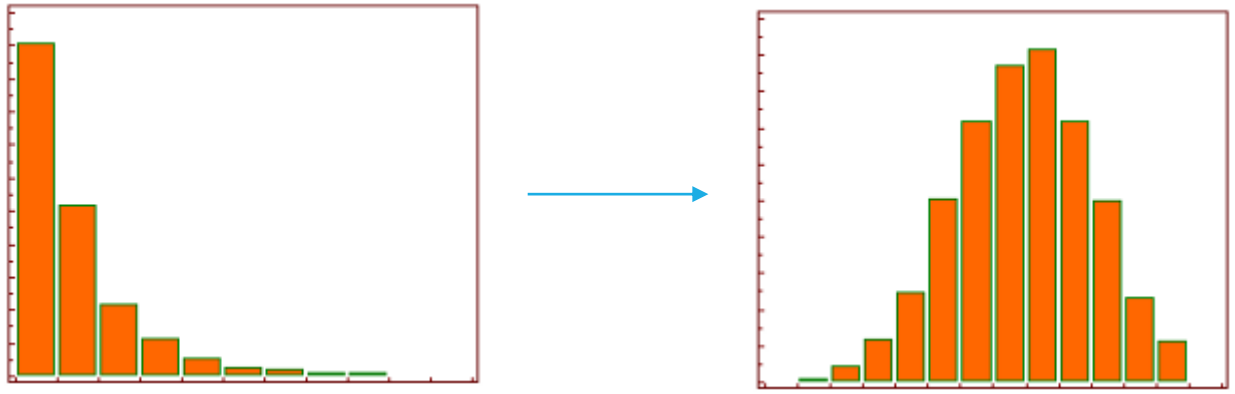
Ridge plot of variable distribution



Traditional Methods

Transformations

Financial data is characterized by **highly right tailed distributions**. Traditional parametric methods for anomaly detection often assume **symmetric** and sometimes **normal** distributions.



- Log transformation
- Box-Cox transformation: $y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & if\ \lambda \neq 0, \\ \ln y_i & if\ \lambda = 0 \end{cases}$

where optimal λ is one which results in best approximation of normal distribution curve

Measures of Variability

- SD – common practice, not robust to outliers, parametric approach - normal distribution
- IQR – common practice, non-parametric, semi-robust
- MAD – robust to outliers, non-parametric but does better with symmetric distributions.
- Double MAD – (*Rosenmai,2013*) like MAD robust to outliers and non-parametric, also takes skewness into consideration.

Measures of Variability

MAD - median absolute deviation

$$MAD(Y) = k * median(|Y_i - median(Y)|)$$

Where k is called the scale factor and is taken to be 1.4826 if data is normally distributed. We take k to be 1 since distribution is unknown.

Outlier threshold is taken to be *alpha* deviations from the median.

Measures of Variability

Double MAD - double median absolute deviation

Upper MAD:

$$\tilde{Y} = median(Y)$$

$$Y^{(u)} = \{y | y \in Y \cap y \geq \tilde{Y}\}$$

$$MAD^{(u)}(Y) = k * median\left(\left|Y_i^{(u)} - \tilde{Y}\right|\right)$$

Where k is called the scaling factor like in original MAD.

Similar for Lower MAD

Outlier threshold is taken to be *alpha* **upper** deviations, and *alpha* **lower** deviations from the median.

Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

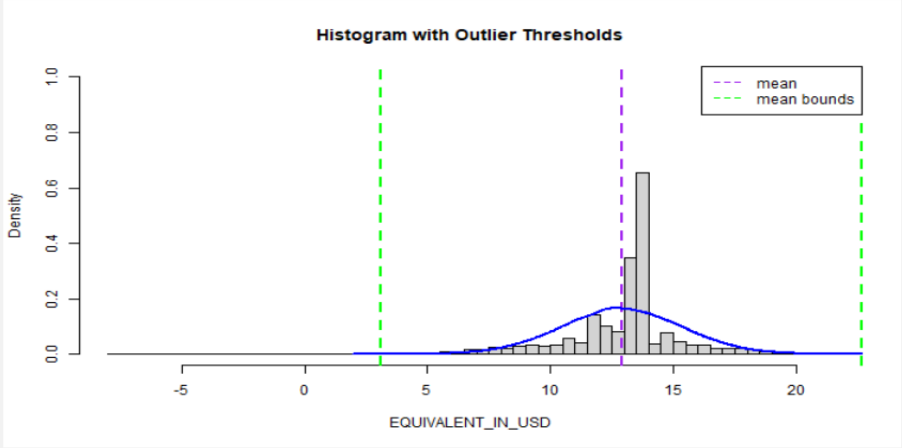
Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability
- (•) Mean/SD
- ( ) Median/MAD
- ( ) Median/DoubleMAD
- ( ) IQR

The Following analysis **does not include NA's and Zero values**



Examine Outliers

Choose Outlier Type
- ( ) Upper
- ( ) Lower
- (•) All

SD

```
[1] "Number of All outliers: 658"
```

Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

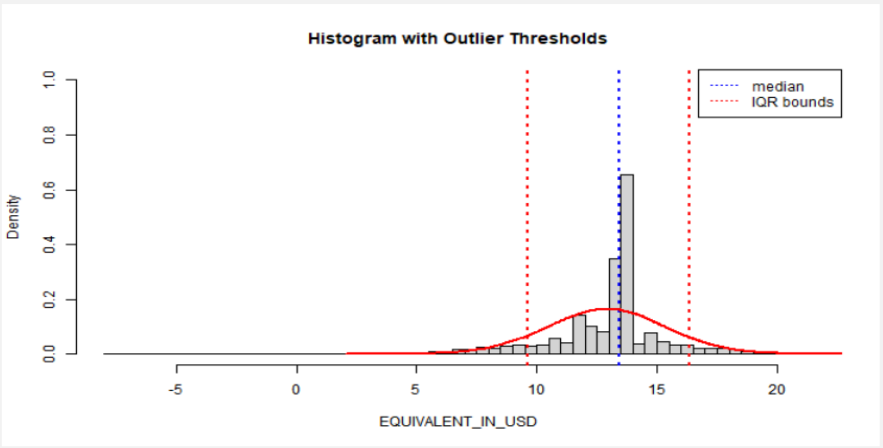
Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability
- ( ) Mean/SD
- ( ) Median/MAD
- ( ) Median/DoubleMAD
- (•) IQR

The Following analysis **does not include NA's and Zero values**



Examine Outliers

Choose Outlier Type
- ( ) Upper
- ( ) Lower
- (•) All

IQR

```
[1] "Number of All outliers 26128"
```

Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

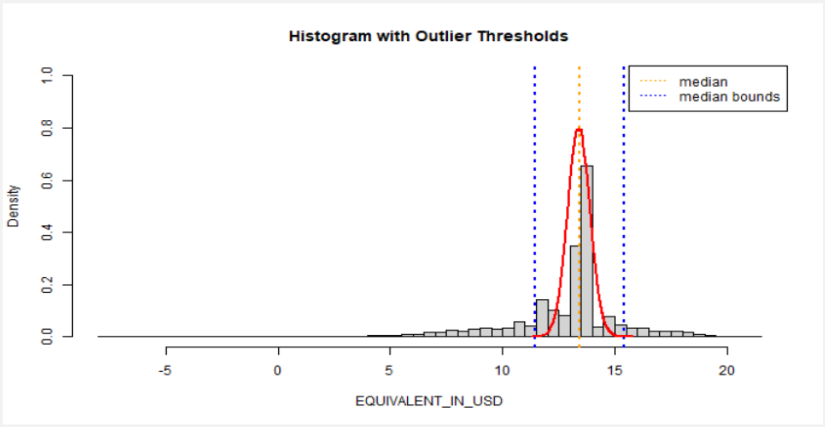
Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability
- ( ) Mean/SD
- (•) Median/MAD
- ( ) Median/DoubleMAD
- ( ) IQR

The Following analysis **does not include NA's and Zero values**



Examine Outliers

Choose Outlier Type
- ( ) Upper
- ( ) Lower
- (•) All

MAD

```
[1] "Number of All outliers 45786"
```

Find Potential Outliers using SD, MAD & IQR

Define the distance factor α. Setting $\alpha = 3$ for Mean/SD method covers 99% of the distribution **assuming the data is normally distributed**.

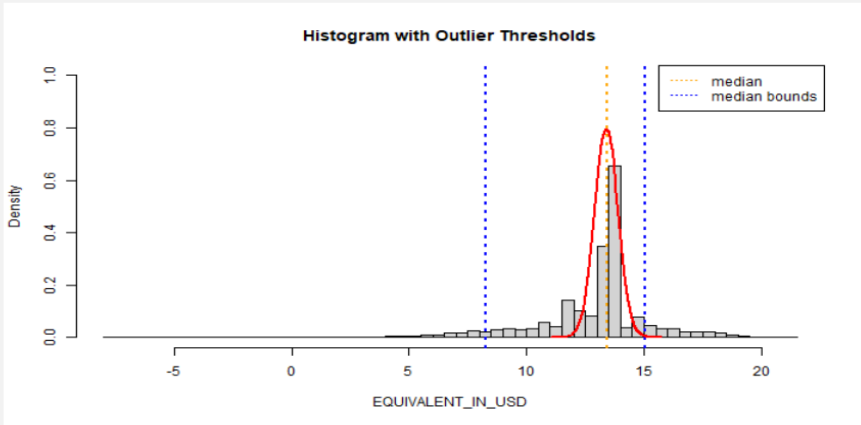
Note that alpha is irrelevent when choosing IQR

Data Transformation: Log

Choose Alpha: 4

Choose Measure of Variability
- ( ) Mean/SD
- ( ) Median/MAD
- (•) Median/DoubleMAD
- ( ) IQR

The Following analysis **does not include NA's and Zero values**



Examine Outliers

Choose Outlier Type
- ( ) Upper
- ( ) Lower
- (•) All

Double MAD

```
[1] "Number of All outliers 28205"
```

Non-Parametric Univariate and Multivariate Methods

Non-Parametric Methods

Applying **transformations** is not always enough to meet the **symmetry** or other **parametric assumptions** required for traditional anomaly detection methods.

A **distribution-free** test for outliers in data drawn from an **unknown data generating process** may give more reliable results.

Non-Parametric Methods

We offer two such methods:

1. Bootlier Plot and Bootlier Test:
   - "Bootlier-Plot – Bootstrap Based Outlier Detection Plot" *Kesar Singh and Minge Xie, 2003*
   - "A Distribution-free Test for Outliers, Discussion Paper Deutsche Bundesbank" *Bertrand Candelon and Norbert Metiu, 2013*
2. Isolation Forest:
   - "Isolation Forest" *Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou*

Bootlier Plot - *Singh and Xie, 2003*

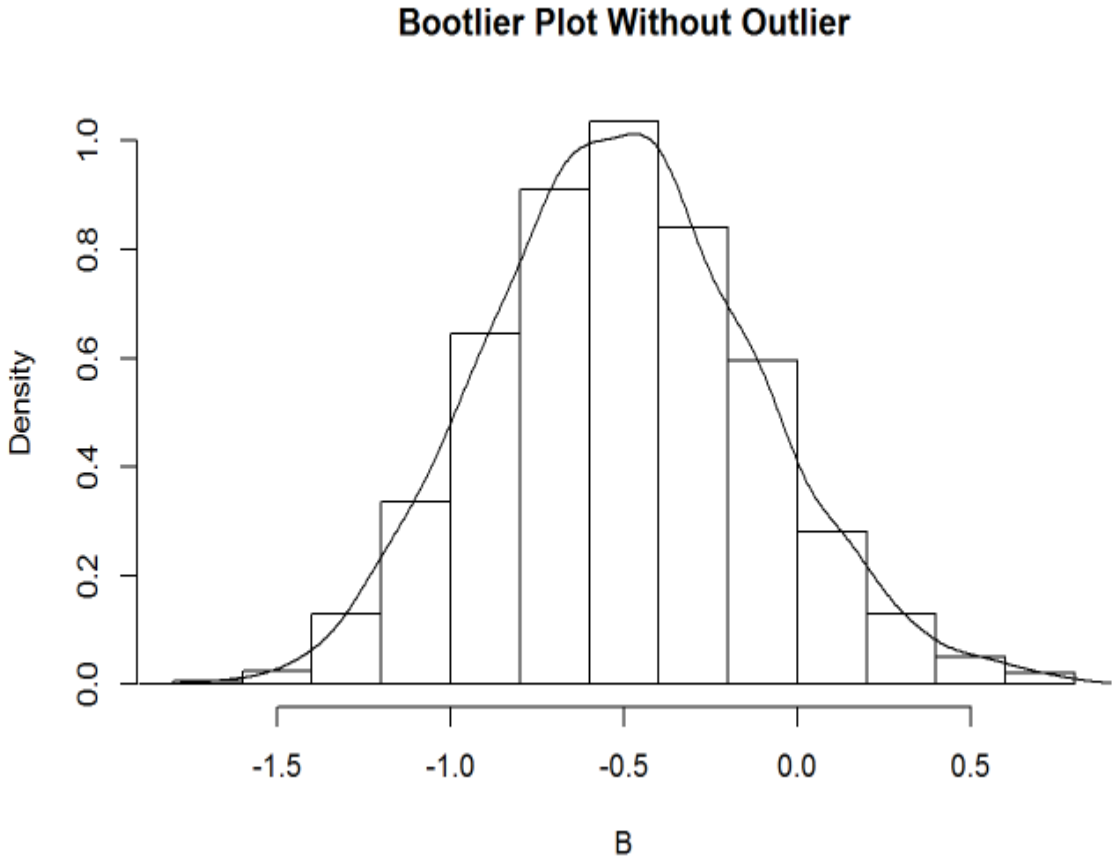
- Method is based on bootstrapping.
- When an outlier exists in a dataset, some bootstrap samples will contain the outlier while others will not.
- Presence of an outlier is expected to cause a significant increase or decrease in the bootstrap mean, and make the bootstrap distribution of the sample mean a mixture distribution.
- We expect the histogram of the sample mean to be multimodal.
- In order to make the bootstrap histogram more sensitive to a potential outlier, the chosen bootstrap statistic is the **"mean – trimmed mean".** Where the trimmed mean is the mean of the bootstrap sample after trimming *k* observations from each side of the sorted sample.

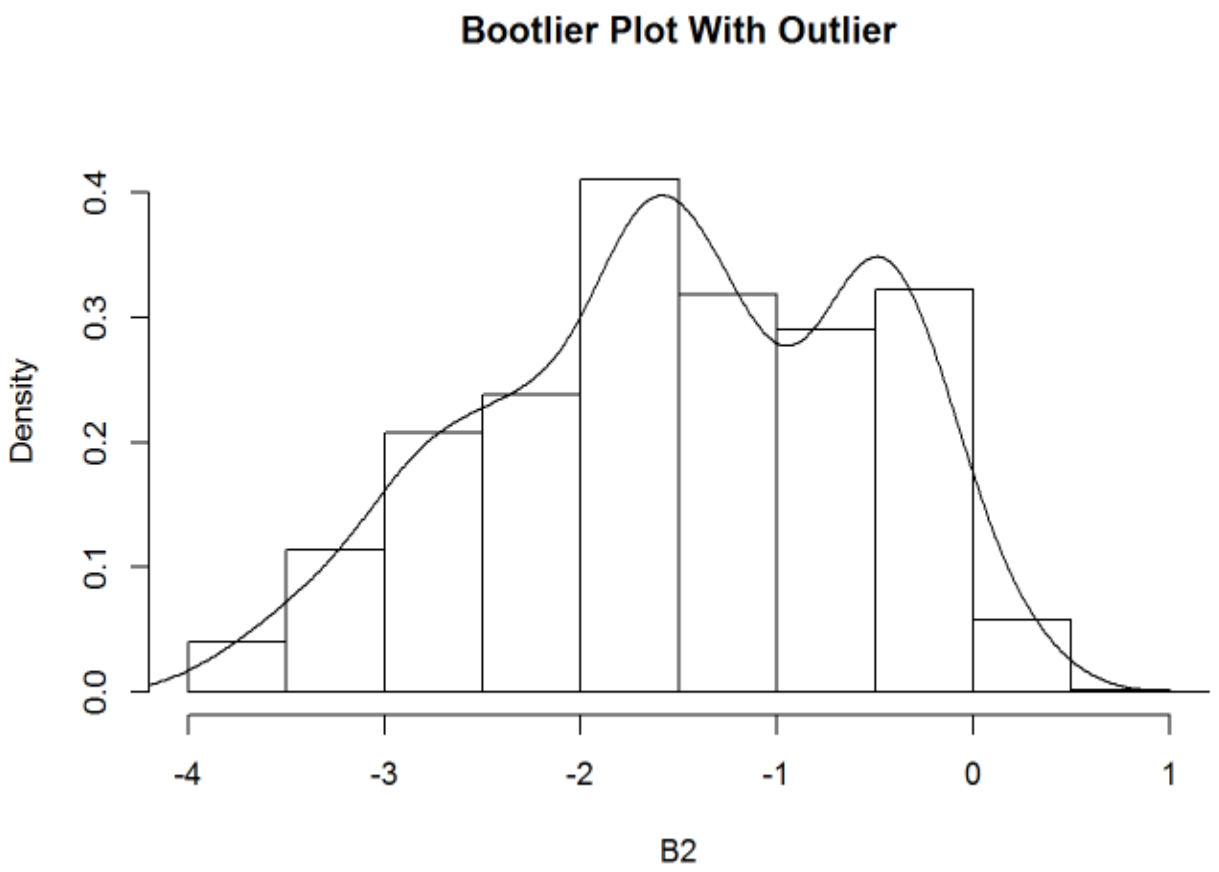
Mean-trimmed mean statistic:

$$T(Y^*) = \frac{1}{n}\sum_1^n Y_i^* - \frac{1}{n-2k}\sum_{k+1}^{n-k} Y_{(i)}^*$$

where $Y_1^*, Y_2^*, \dots, Y_n^*$ denote bootstrap draws and $Y_{(1)}^*$'s be the corresponding order statistics

Bootlier Plot





Bootlier Plot - *Singh and Xie, 2003*

Bootlier Plot for large sample with numerous outlier candidates:

- Probability of having a bootstrap sample of size *n*, free from potential outliers is very small and we may not see a clear bump in Bootlier plot.
- Trick is to reduce bootstrap sample size to a fraction of *n*, i.e., *[αn]*, $\alpha \in (0,1]$
- **Practical recommendation:** look at several Bootlier plots at different bootstrap sample sizes – if any one is found bumpy that would be indicative of the presence of outliers.

Bootlier Test - *Candelon and Metiu, 2013*

Identification of outliers based on Bootlier Plot

Bootlier Plot tells us whether there are outliers or not – how do we identify the outliers themselves?

Candelon and Metiu, 2013 address this in the Deutsche Bundesbank Discussion Paper – "A distribution-free test for outliers".

- Two step process:
  1. **Test for multimodality:** **H0** – Bootlier plot has precisely one mode (and no local minimum), **H1** – Bootlier plot has more than one mode. Test hypothesis using:
     "Bootlier Test" - Bootlier plot coupled with distribution-free test for multimodality proposed by Silverman (1981)
  2. **Identify Outliers:** **1)**Build subsamples by sequentially canceling observations from the tails of the original sample ordered in ascending order. **2)**Perform Bootlier test on each ordered subsample until the null hypothesis of unimodality cannot be rejected for a particular subset of observations. **3)**Data points not contained in this subset are the outliers.

Anomaly Detection App Shir Kamenetsky — 29-09-2021

Data Upload and Preparation | Examining the Variables | Parametric Outlier Detection | Non-Parametric Outlier Detection: iForest | Non-Parametric Outlier Detection: Bootlier Plot

Parameters

Choose Bootstrap Size

1000

Choose Bootstrap Sample Size

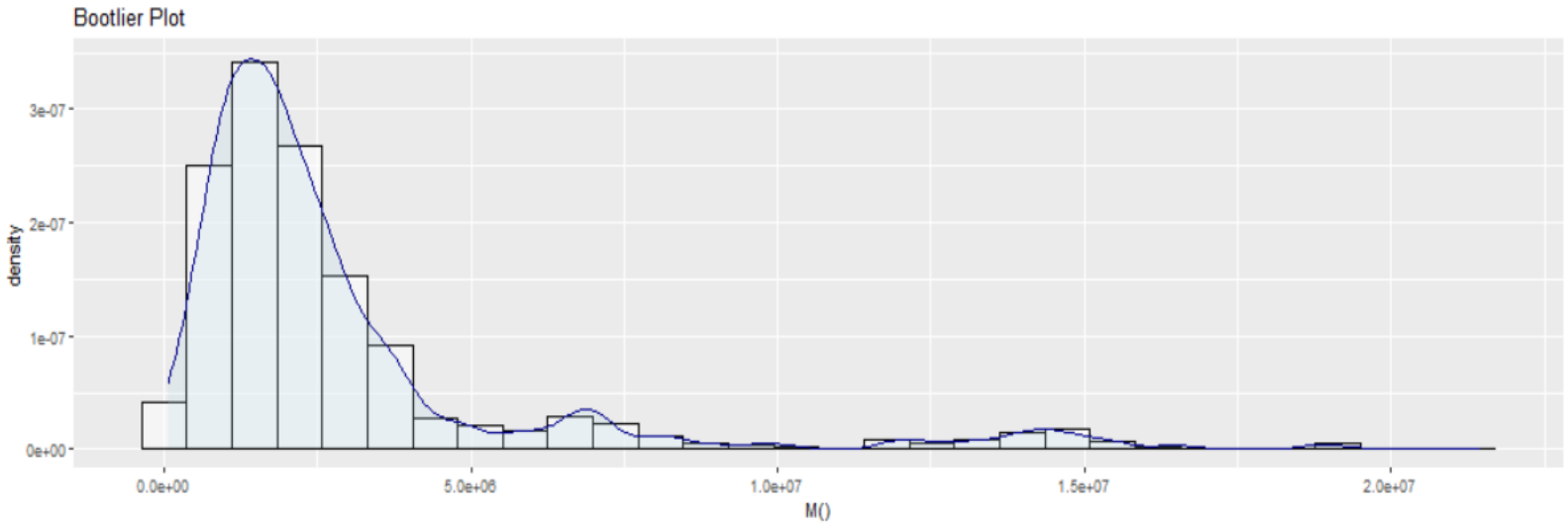
100

Choose Trim Amount

2

Run

Bootlier Plot



Bootlier Outliers

```
[1] 0.00044999
```

Anomaly Detection App Shir Kamenetsky — 29-09-2021

Data Upload and Preparation | Examining the Variables | Parametric Outlier Detection | Non-Parametric Outlier Detection: iForest | Non-Parametric Outlier Detection: Bootlier Plot

Parameters

Choose Bootstrap Size

1000

Choose Bootstrap Sample Size

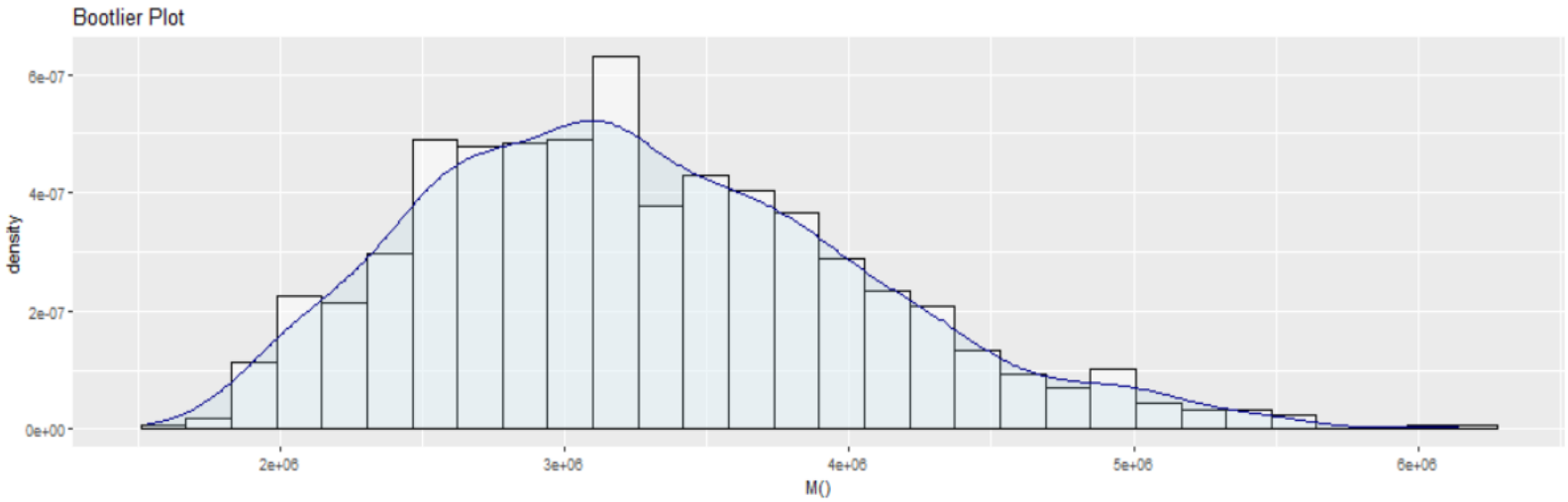
2000

Choose Trim Amount

40

Run

Bootlier Plot



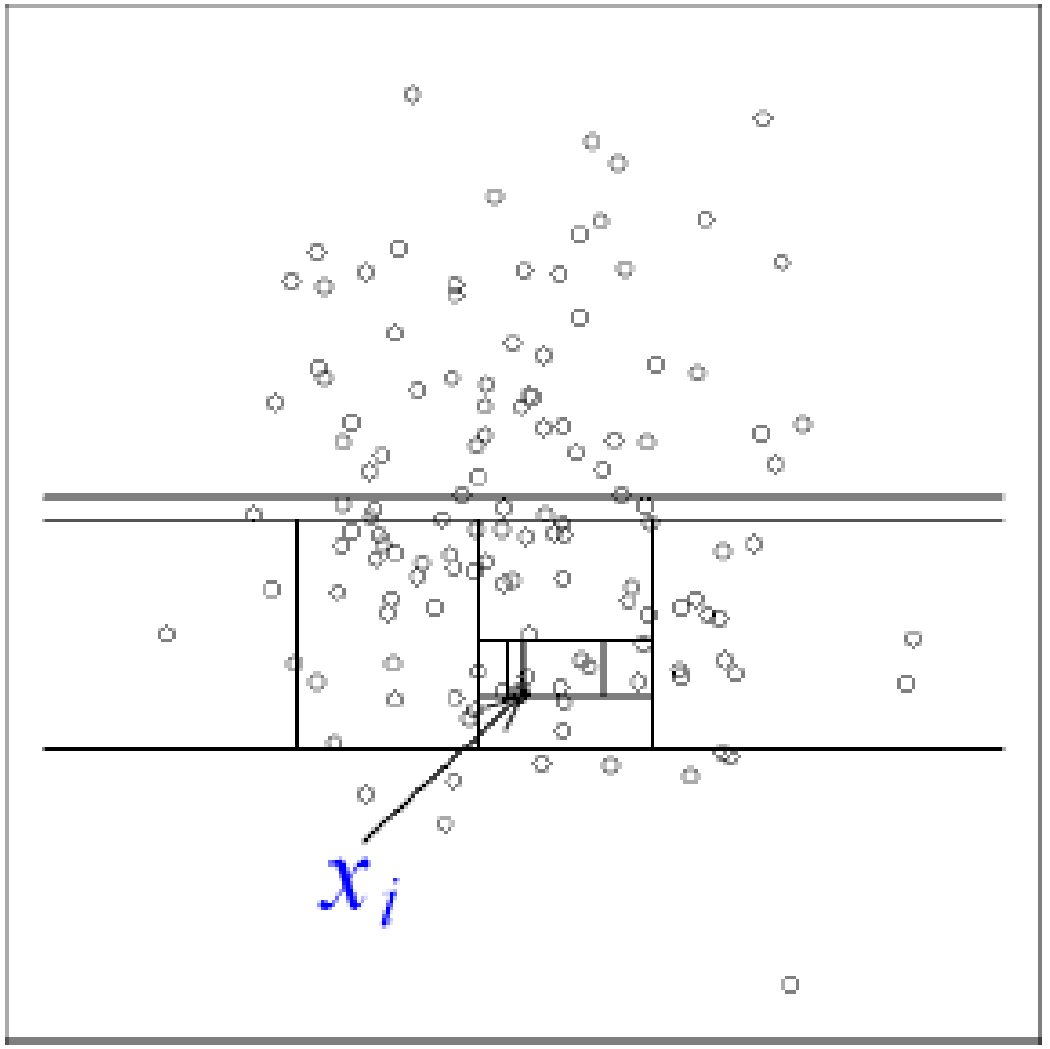
Bootlier Outliers

NULL

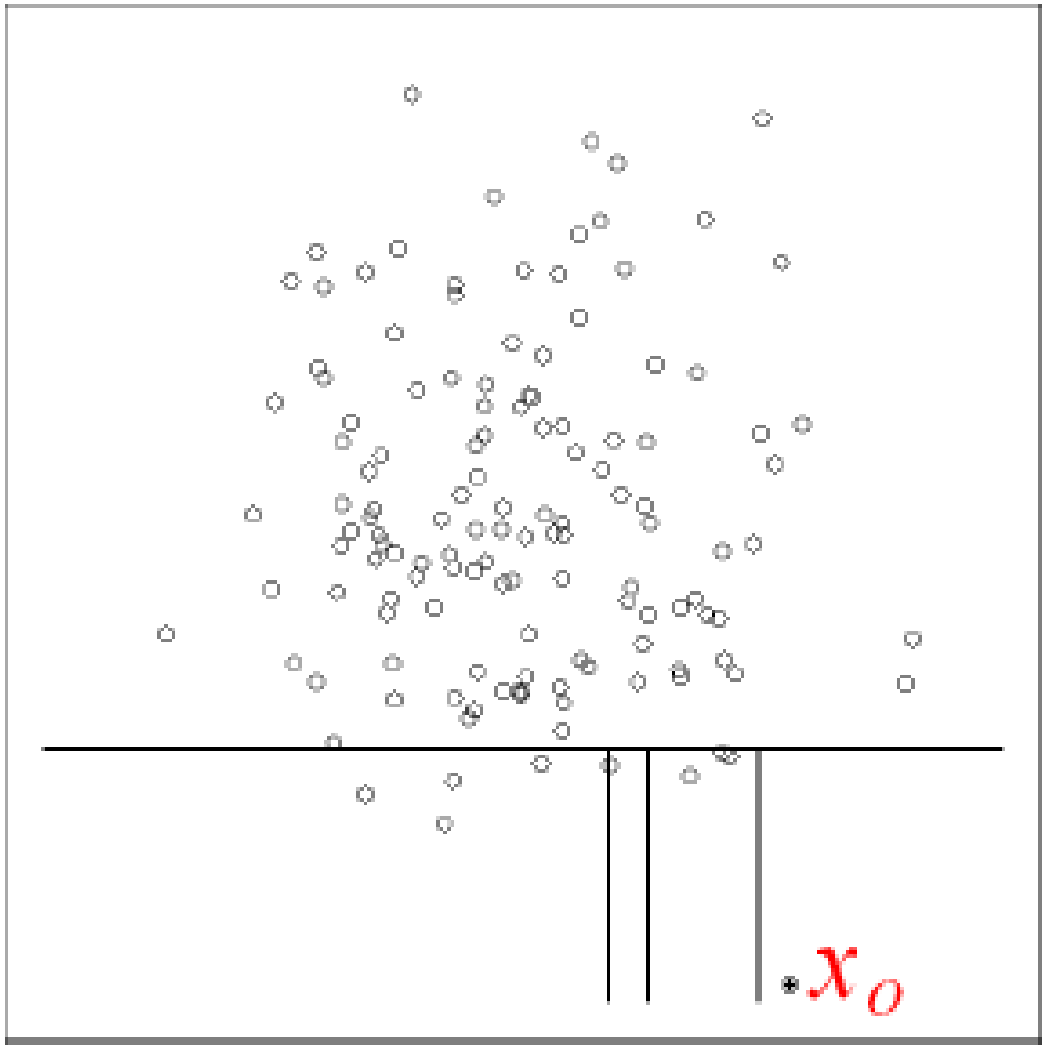
Isolation Forest - *Liu, Ting, and Zhou*

Basic idea is to **isolate anomalies** rather than profiling normal instances. Anomalies are **"few"** and **"different"** making them more susceptible to isolation than normal points. Isolation is conducted via construction of **tree structure**.

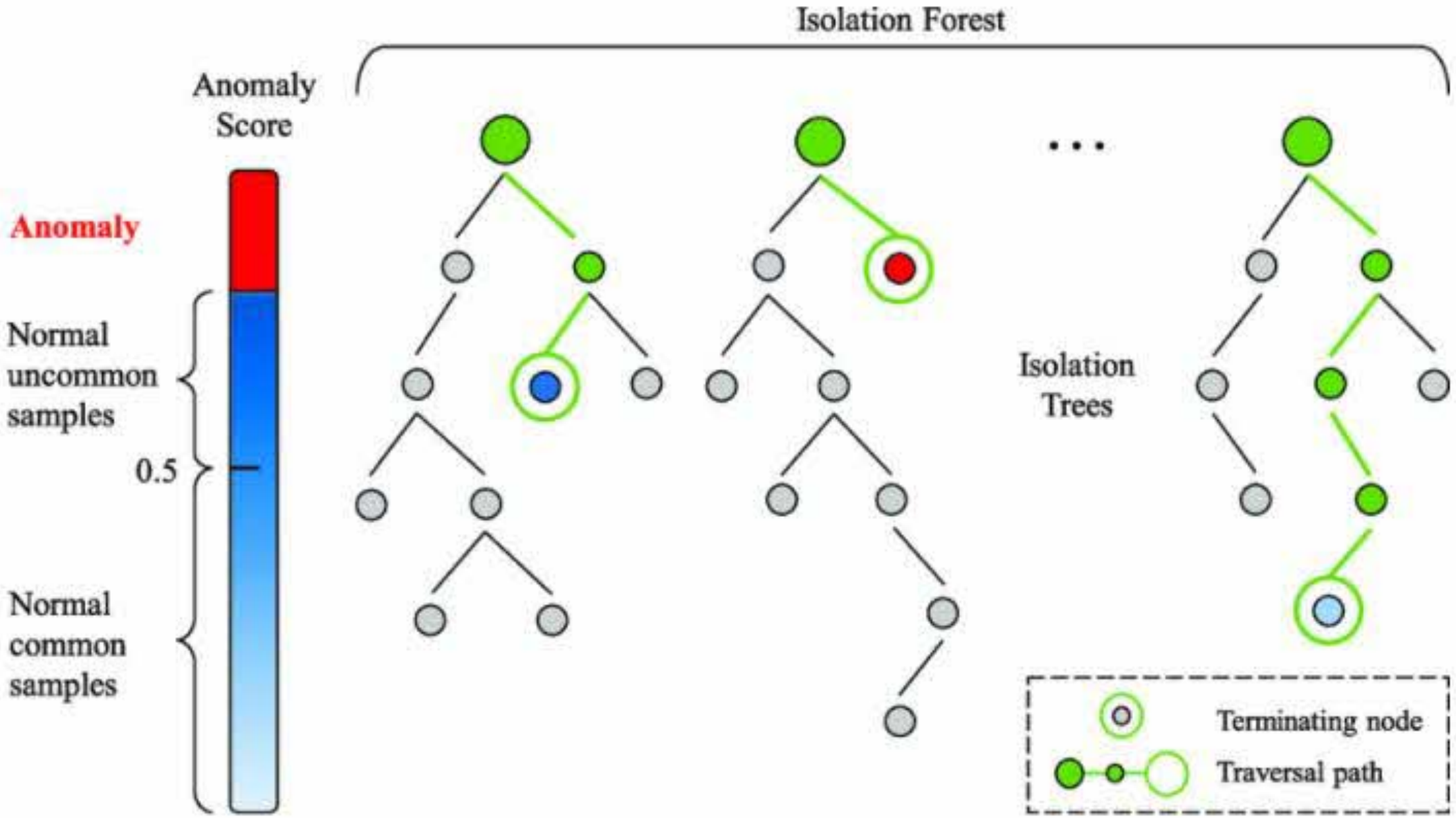
- **iTree (Isolation Tree)** – proper binary tree.
  - Partitions are generated by randomly selecting an attribute and then randomly selecting a split value between the maximum and minimum values of selected attribute.
  - This is done iteratively until each observation is isolated to its own node.
  - **Path length** is equivalent to number of partitions required to isolate a point, or in tree structure the number of edges from root node to external terminating node.
- **Iforest (Isolation Forest)** –
  - Builds ensemble of iTrees on sub-samples of data.
  - Calculate **Anomaly Score**: Aggregate iTree paths for each observation.
  - Anomalies are those instances which have short path lengths – or **anomaly scores close to 1**.



(a) Isolating x_i



(b) Isolating x_o


Isolation Forest
Anomaly Score
Anomaly
Normal uncommon samples
0.5
Normal common samples
Isolation Trees
Terminating node
Traversal path

Anomaly Detection App Shir Kamenetsky — 29-09-2021

Data Upload and Preparation | Examining the Variables | Parametric Outlier Detection | Non-Parametric Outlier Detection: iForest | Non-Parametric Outlier Detection: Bootlier Plot

iForest Anomalies

Number of trees: 100. Sample size: 256. Scores close to 1 are considered anomalies.

Choose Variables

log_EQUIVALENT_IN_USD
UPI_UNIQUE_PRODUCT_IDENTIFIER
UNIFORM_EXCHANGE_RATE_FIX
UNIFORM_EXCHANGE_RATE_BASIS

Anomaly Score Threshold

0.75

Run

Table of Anomaly Scores:

Show 10 entries Search:

	log_EQUIVALENT_IN_USD	UPI_UNIQUE_PRODUCT_IDENTIFIER	UNIFORM_EXCHANGE_RATE_FIX	UNIFORM_EXCHANGE_RATE_BASIS	anomaly_score	anomaly
	All	All	All	All	All	All
146210	0.753669378307685	FXSWAP:Spot	0.3661	SEK/ILS	0.769114463711078	outlier
180035	0.702376082929321	FXSWAP:Spot	0.3661	SEK/ILS	0.769114463711078	outlier
225674	-1.23866404703587	Foreign Exchange:Spot	0.5076	TRY/ILS	0.767554930124541	outlier
238771	0.146471901900257	Foreign Exchange:Spot	0.5106	TRY/ILS	0.767554930124541	outlier
241383	-0.562183920266133	Foreign Exchange:Spot	0.5146	TRY/ILS	0.767554930124541	outlier

Showing 1 to 10 of 290,125 entries

Quantile Table of Anomaly Scores:

```
      50%       55%       60%       65%       70%       75%       80%       85%
0.5827973 0.5831917 0.5843767 0.5859603 0.5875483 0.5903376 0.5955529 0.6061221
      90%       95%      100%
0.6189693 0.6424364 0.7691145
```

```
[1] "Number of outliers: 113"
```

Thank You!