EVERGREEN

TITLE

SUBTITLE

YOUR NAME HERE

YOUR TITLE HERE

- Experience

CALL OUT BOX – USE THIS FOR A QUOTE OR PERSONAL HIGHLIGHT

Certification Badges can go above this text line / this text line can be used for further certification call outs

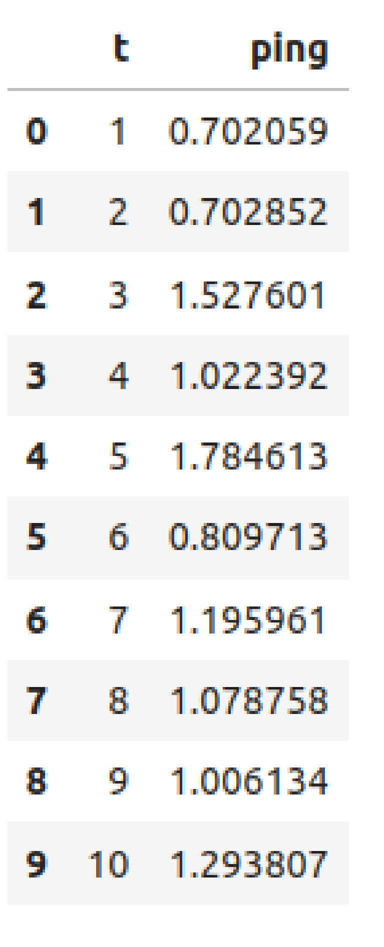
SECTION TITLE

Outlines

- Objective
- About Datasets
- Data Insights or Exploratory Data analysis
- Signal Decomposition
- Anomalies Detection
    - STL
    - Isolation Forest
    - IQR Method
- Conclusion
- Future Plans

About Datasets

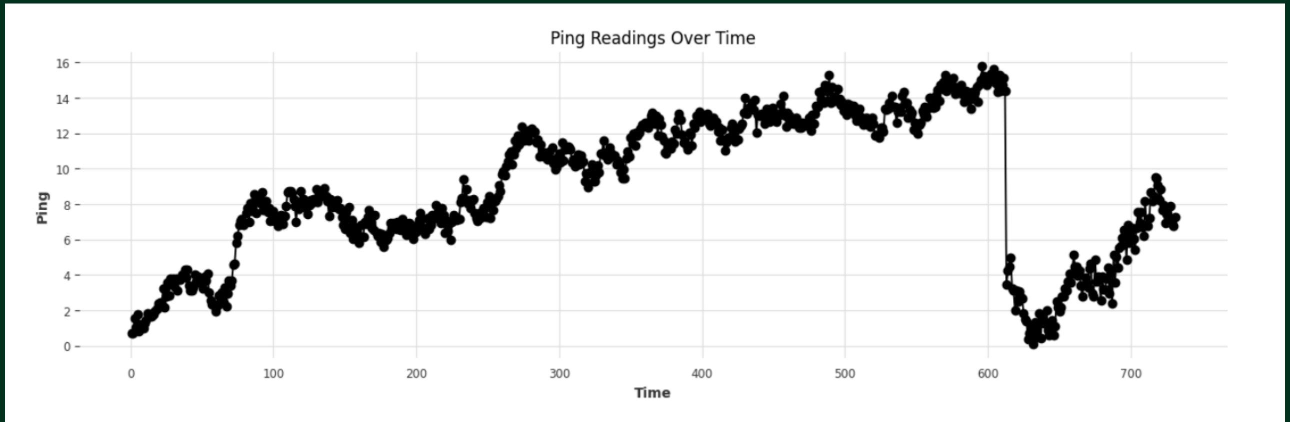
- The given Datasets contain two columns T(time) and ping (ping readings) and size of total datasets is 731 records.
- Since no unit representation is defined in the dataset, the following assumptions were made
  - The t column represents integer data type and is considered as time in seconds.
  - The ping column represents the corresponding ping value at a given second. Ping is measured in milliseconds, but here we have assumed it as a float datatype.



	t	ping
0	1	0.702059
1	2	0.702852
2	3	1.527601
3	4	1.022392
4	5	1.784613
5	6	0.809713
6	7	1.195961
7	8	1.078758
8	9	1.006134
9	10	1.293807

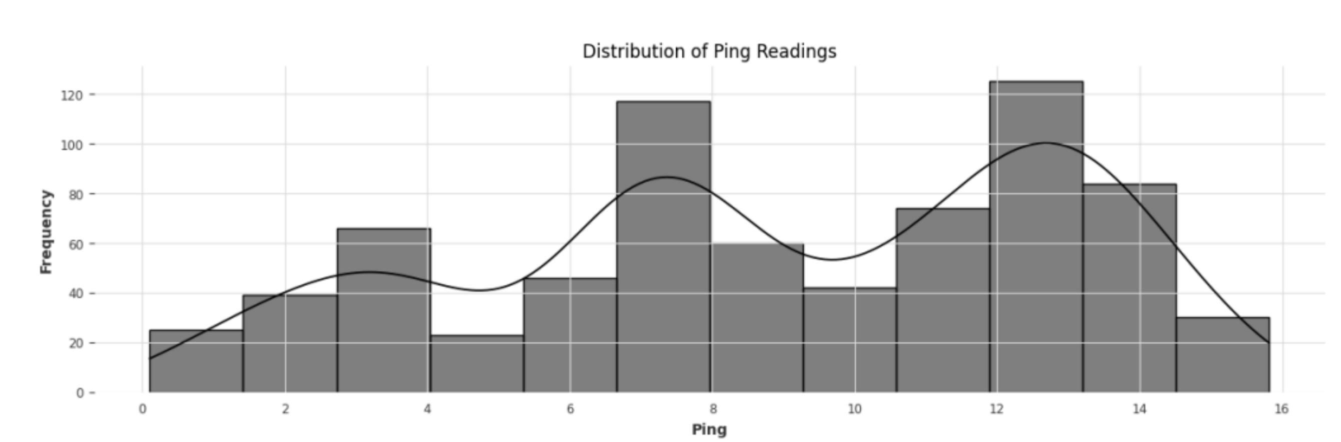
Data Insight : Line Plot

- Given image is represent the Line plot of our datasets.
- X-axis represent the time and y-axis represent the ping readings.
- The line plot shows the variations of ping readings over time and seems to be some fluctuation with no clear trend or periodicity at first glance.



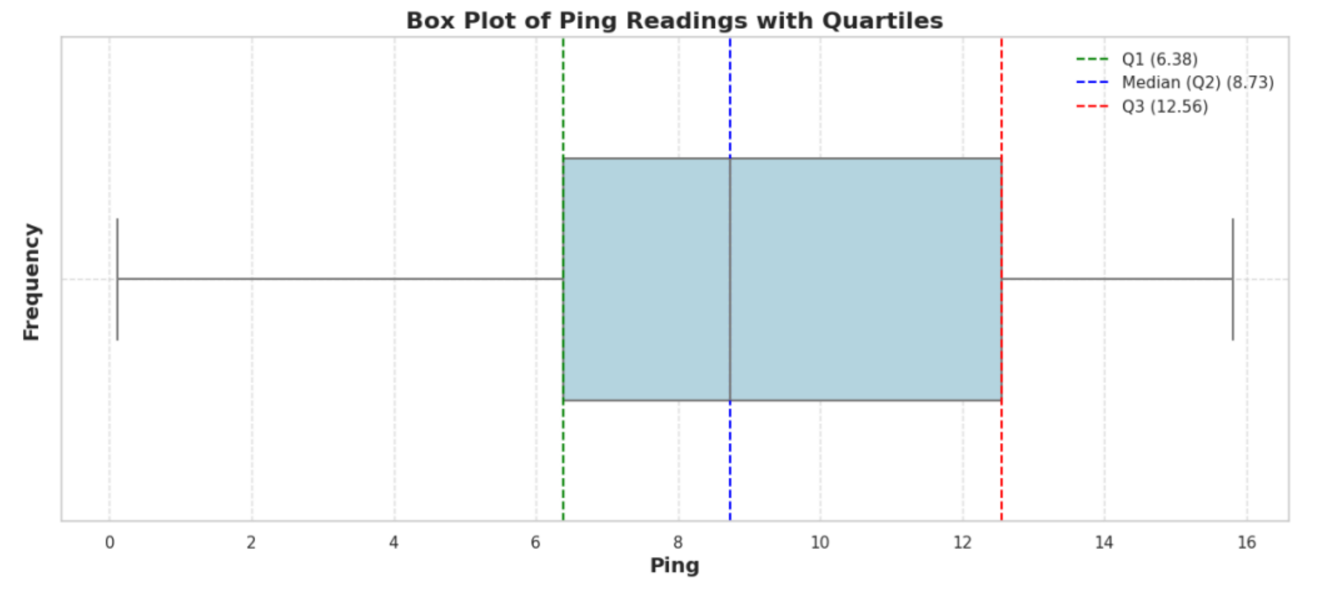
Data Insight : Histogram

- The histogram displays the frequency distribution of ping readings.
- The x-axis represents the range of ping values, while the y-axis shows the frequency (or count) of ping readings within each bin. The height of each bar indicates how many ping values fall within the corresponding range.
- The histogram combined with the KDE curve reveals the overall shape of the distribution. For example, if the KDE curve is unimodal (has a single peak), it suggests that the ping values are concentrated around a central value.



Data Insight : Boxplots

- Given plot is Boxplots which visualize the distribution of ping readings, including the median, quartiles and potential outliers where
  - Q1 Line (Green): Indicates the first quartile (25th percentile). 25% of the ping values are below this line.
  - Median Line (Blue): Shows the median (50th percentile), representing the central value of the ping readings.
  - Q3 Line (Red): Represents the third quartile (75th percentile). 75% of the ping values are below this line.
  - Outliers: Points beyond the whiskers are considered outliers



Signal Decomposition using STL

Seasonal-Trend decomposition using LOESS (STL) is a robust method of time series decomposition often used in economic and environmental analyses. The STL method uses locally fitted regression models to decompose a time series into trend, seasonal, and remainder components. we can apply STL to any dataset, but meaningful results are only returned if a recurring temporal pattern exists in the data

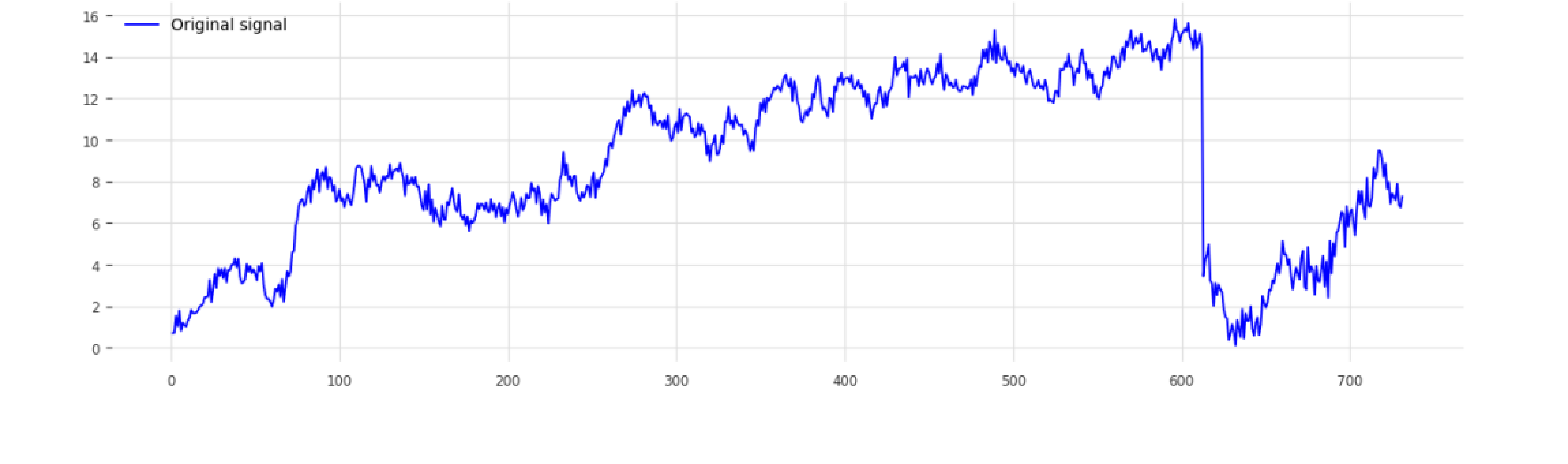
The STL algorithm performs smoothing on the time series using LOESS in two loops; the inner loop iterates between seasonal and trend smoothing and the outer loop minimizes the effect of outliers. During the inner loop, the seasonal component is calculated first and removed to calculate the trend component. The remainder is calculated by subtracting the seasonal and trend components from the time series

The three components of STL analysis relate to the raw time series as follows:

yi = si + ti + ri

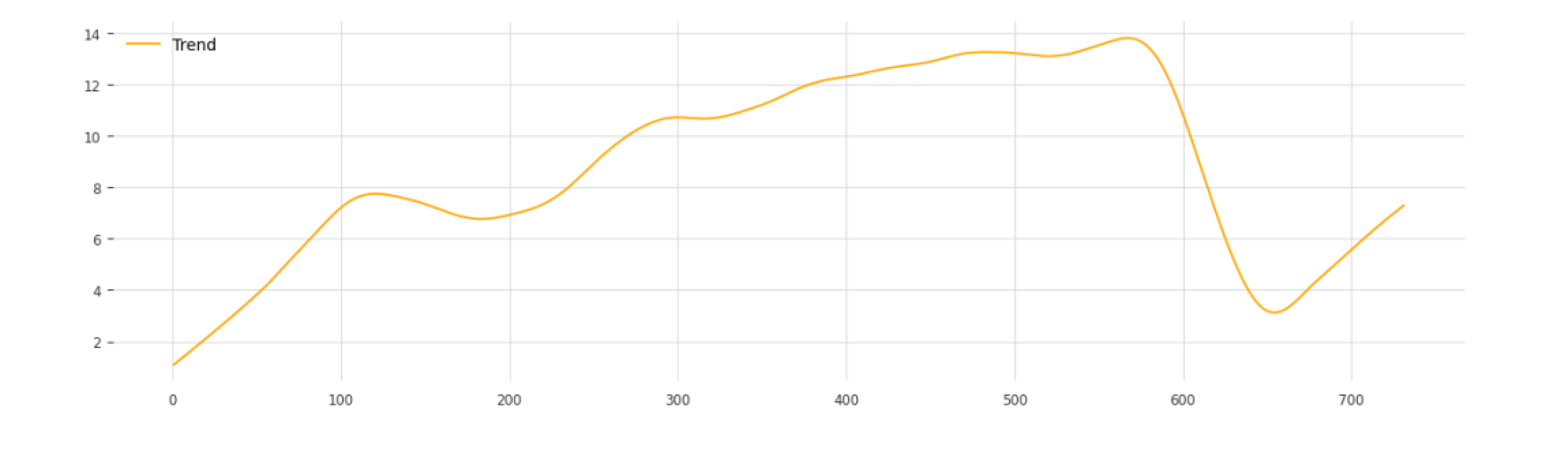
- yi = The value of the time series at point i.
- si = The value of the seasonal component at point i.
- ti = The value of the trend component at point i. • ri = The value of the remainder component at point i

Signal Decomposition : Orignal Series



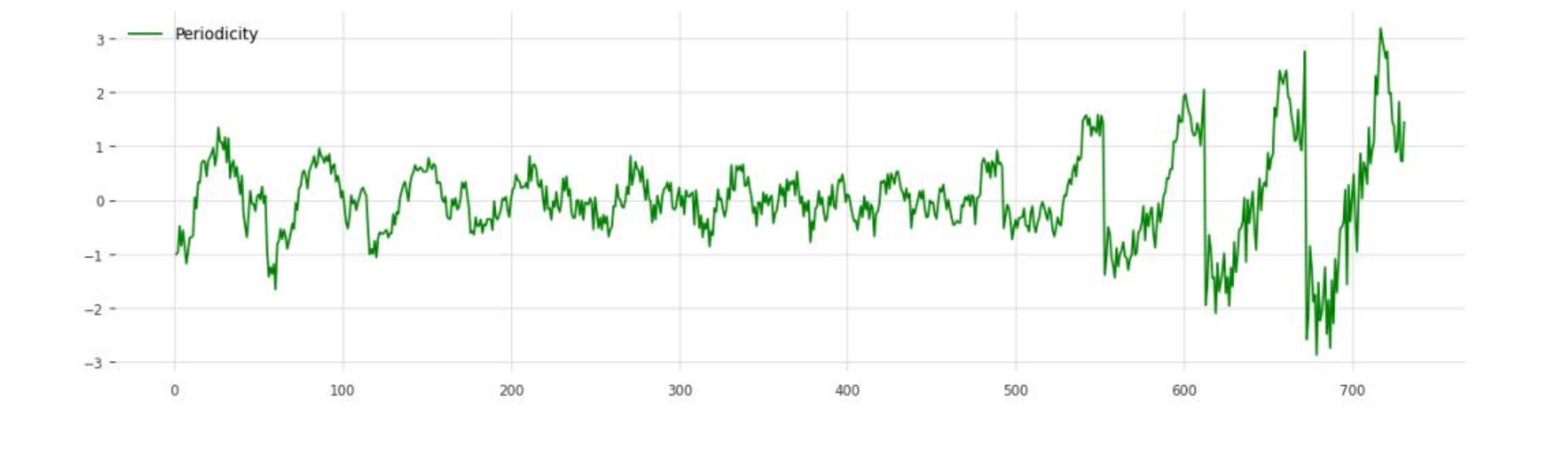
Signal Decomposition : Trend

The orange line represents the trend component extracted from the time series data. This component shows the general direction or long-term movement.



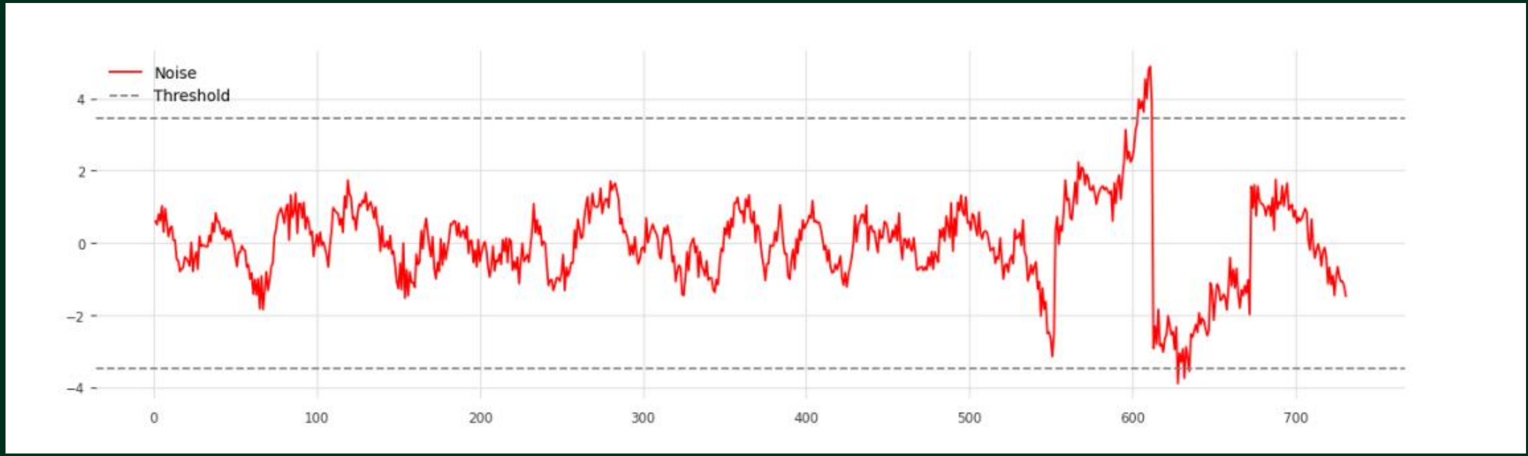
Signal Decomposition : Periodicity

The green line represents the seasonal component extracted from the time series data. This component shows the periodic patterns or cycles that repeat over regular intervals. Given plot shows the recurring cycles or patterns within the data. from the observation data don’t have strong consistent recursive pattern or the data contain have more than one periodicity.



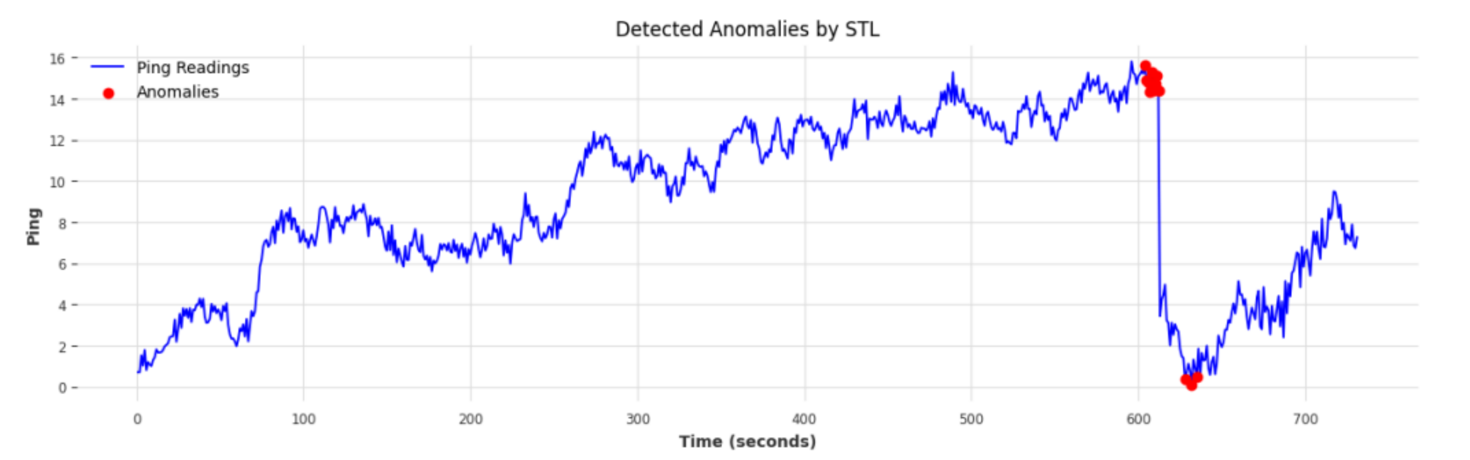
Signal Decomposition : Noise

- Noise (Residual): The plot shows the residual component of the time series, which represents the part of the data that remains after removing the trend and seasonal components. It is plotted in red, indicating the deviations from the expected values.
- Threshold Lines): The horizontal dashed lines represent the upper and lower thresholds. These thresholds are used to detect significant deviations or anomalies in the residuals.
- The regions where the noise consistently crosses the threshold lines are of particular interest as they indicate points where the data deviates significantly from the norm.



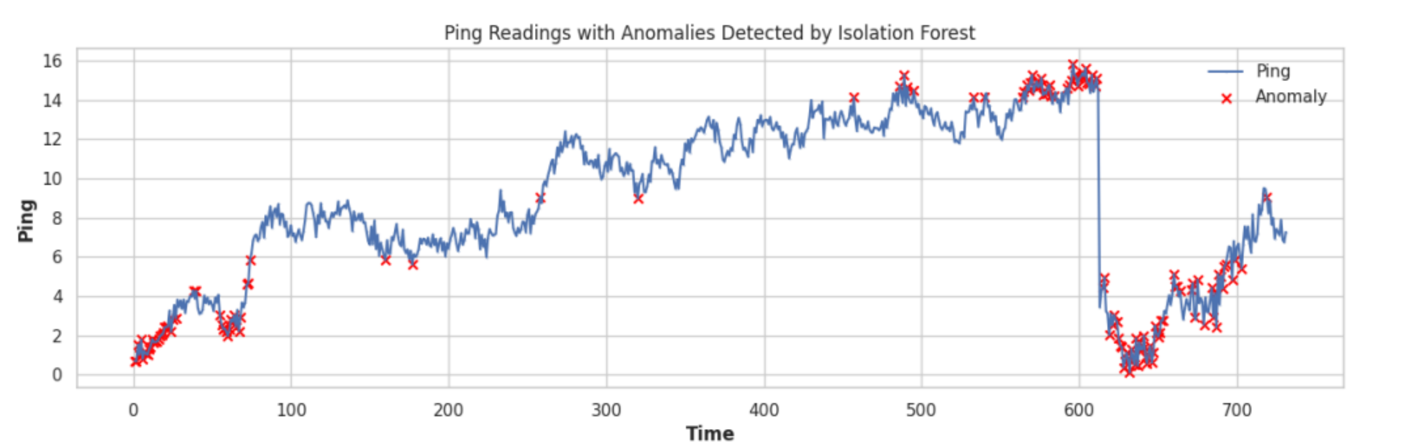
Signal Decomposition : Detected Anomalies

- The blue line represents the original ping readings over time, showing the overall behavior of the data, including any trends and seasonal fluctuations.
- The red scatter points indicate the locations where anomalies have been detected. These points represent deviations from the expected pattern, as identified by the anomaly detection using STL.
- These deviations could be due to unusual spikes or drops in latency

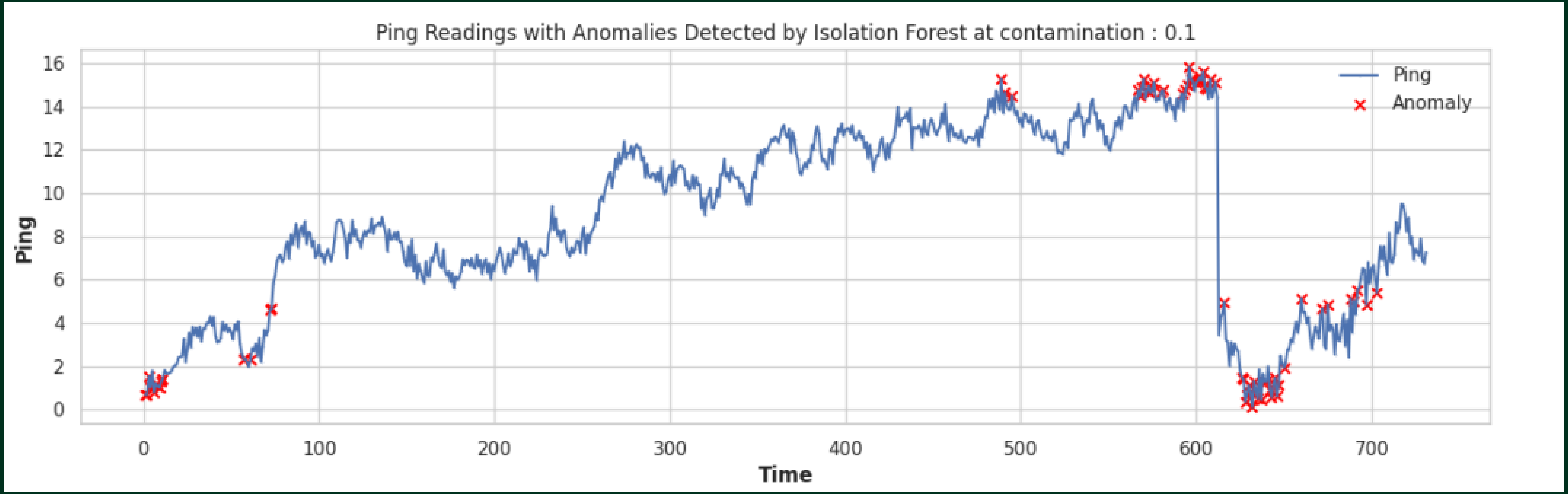


Anomaly Detection : Isolation Forest

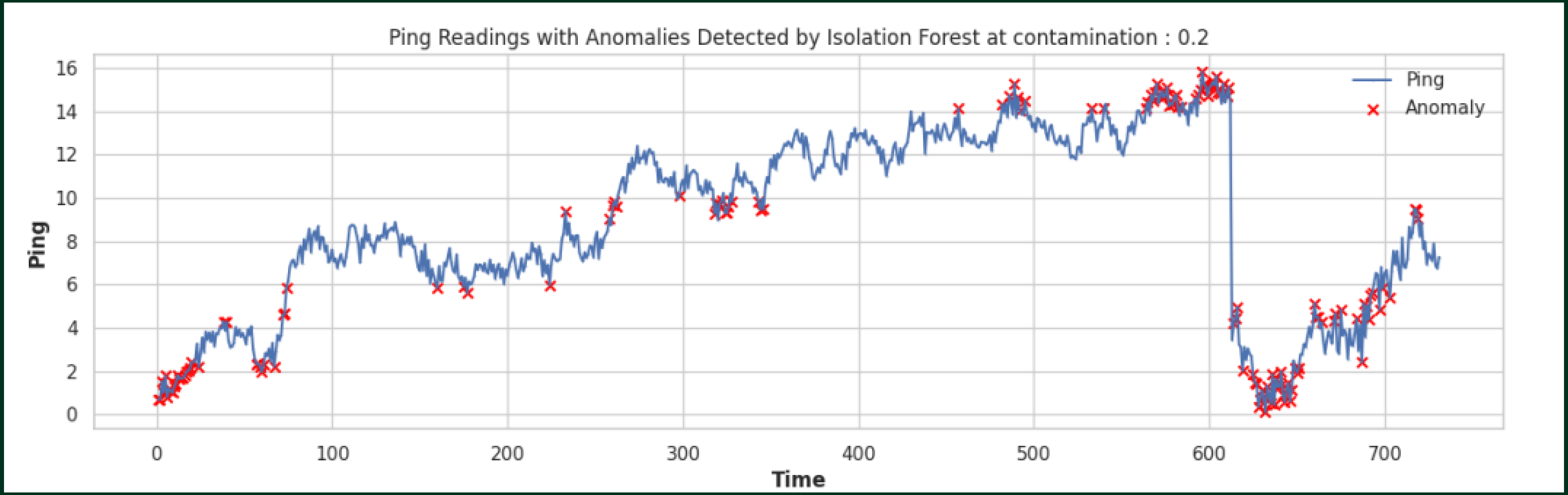
- Isolation Forest is an anomaly detection algorithm based on the concept of isolating observations.
- The algorithm isolates anomalies instead of profiling normal data points. It constructs a number of random trees (decision trees) where each tree partitions the data randomly.
- Anomalies are more likely to be isolated quickly because they are different from the majority of the data.
- Each observation’s path length in the trees is measured. Anomalies are expected to have shorter path lengths because they are isolated earlier in the trees.
- Contamination Parameters :This parameter is used to estimate the proportion of outliers in the data. It helps in adjusting the threshold for classifying an observation as an anomaly



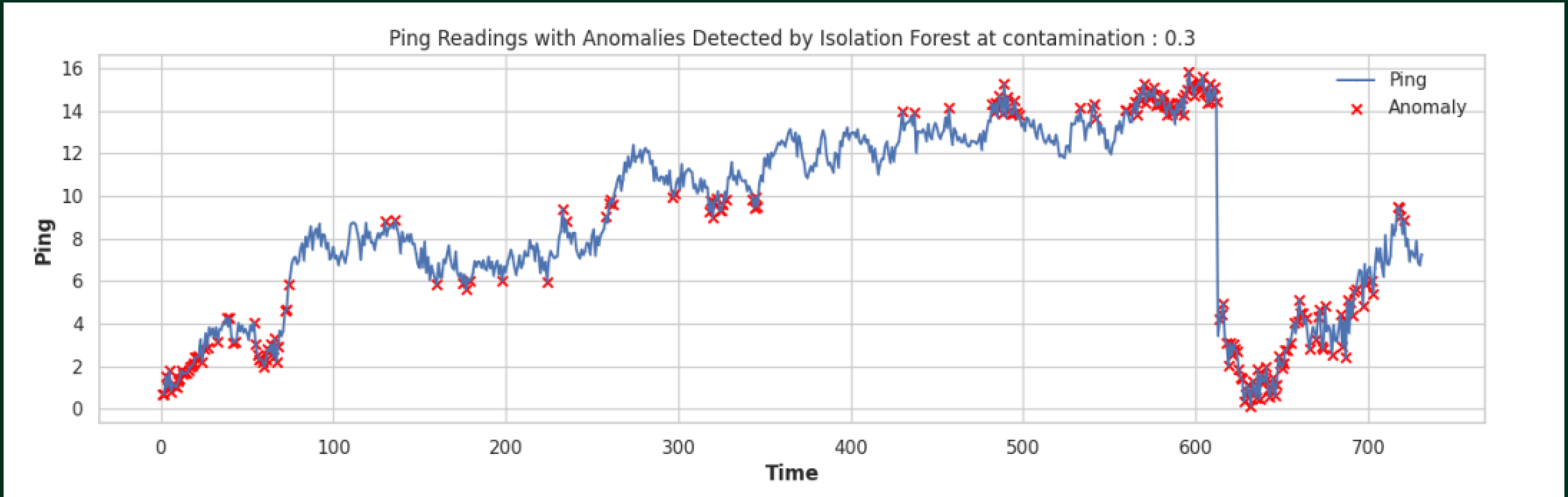
Isolation Forest at contamination [0.1]



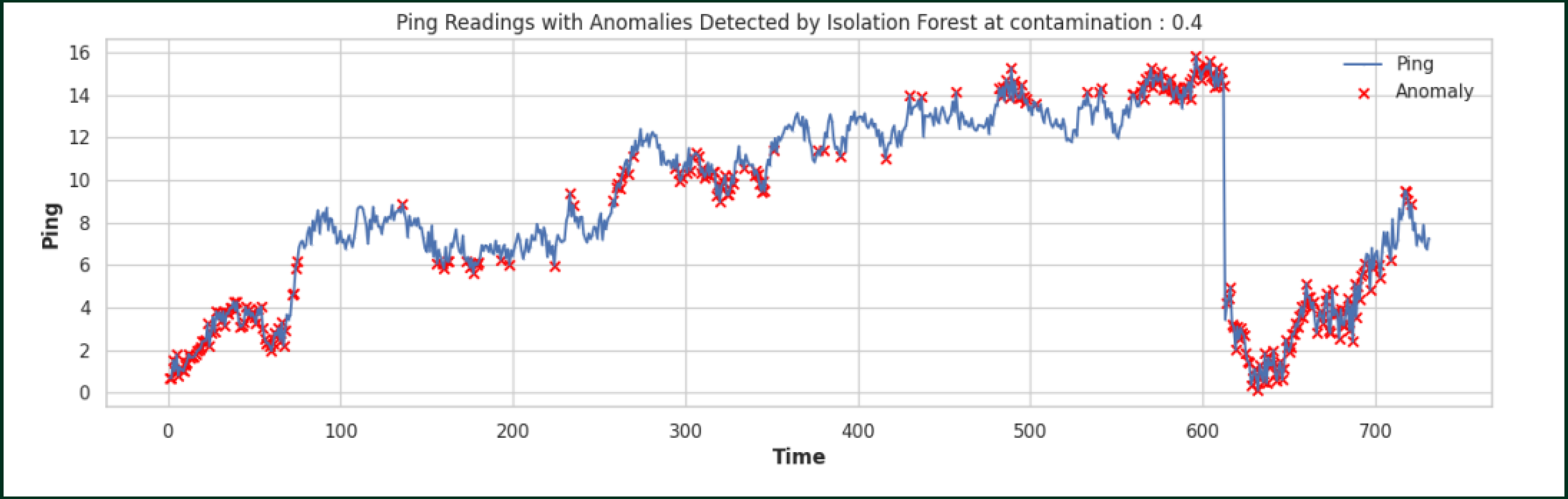
Isolation Forest at contamination [0.2]



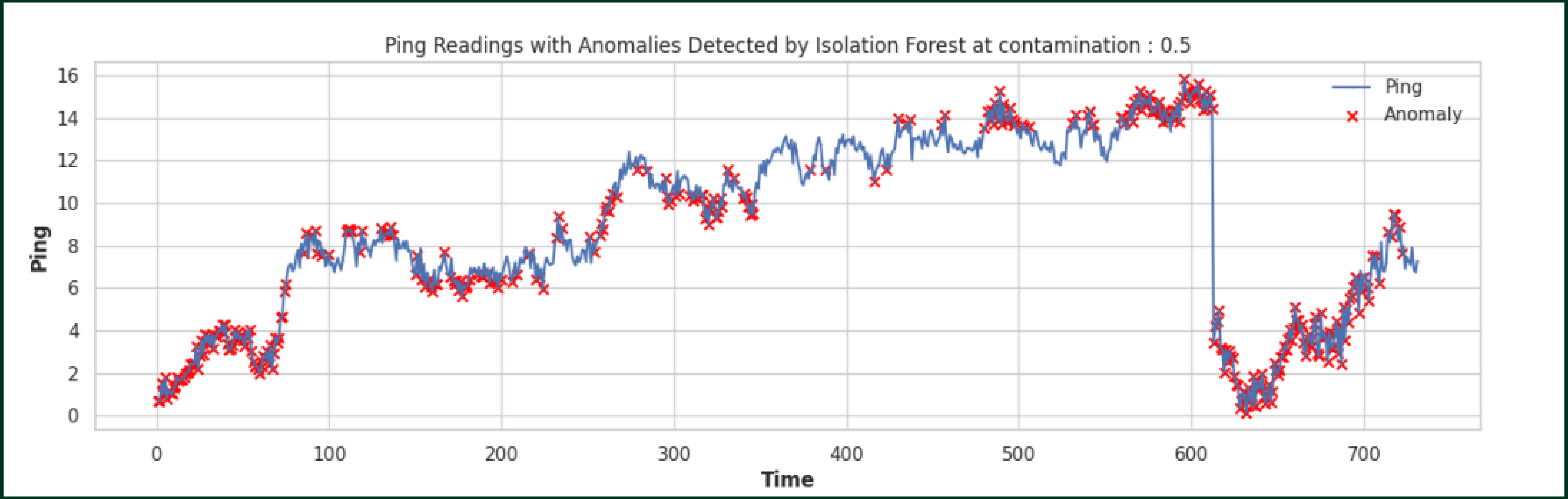
Isolation Forest at Contamination [0.3]



Isolation Forest at contamination [0.4]

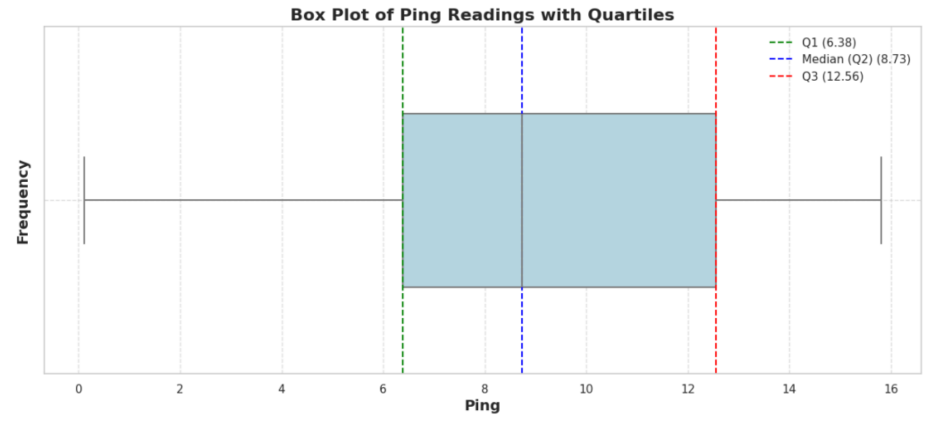


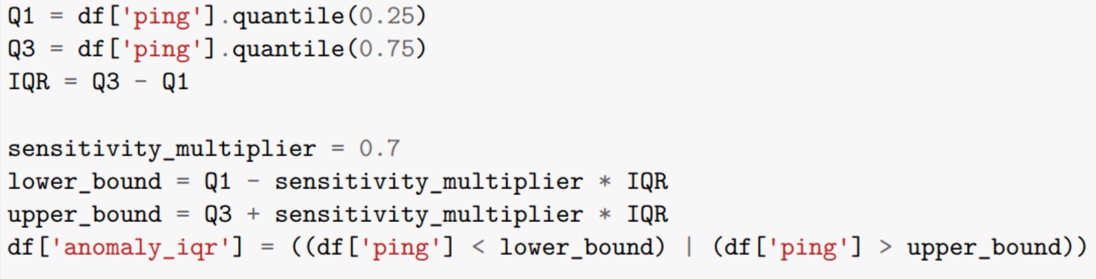
Isolation Forest at contamination [0.5]



IQR Method : Anomaly Detection

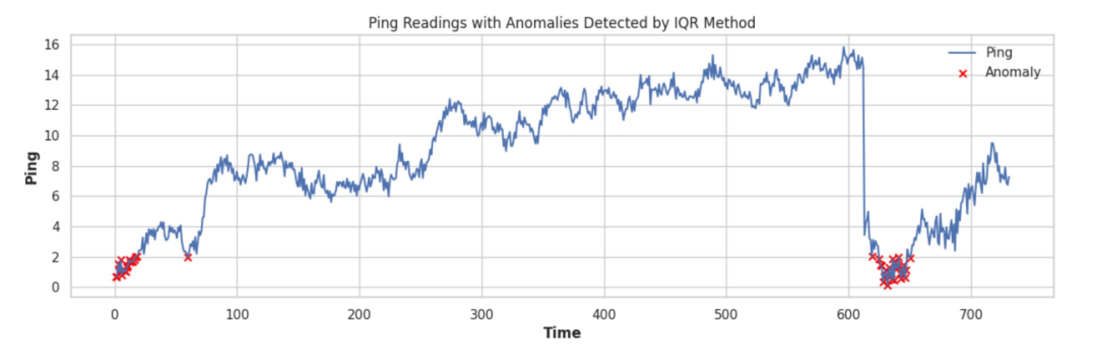
- Q1 and Q3 divide the data into four equal parts. Q1 is the value below which 25% of the data falls, and Q3 is the value below which 75% of the data falls.
- The range between Q1 and Q3, which contains the middle 50% of the data. It measures the spread of the central portion of the data.
- Adjusts the width of the anomaly detection bounds. A smaller multiplier results in a narrower range, making it more sensitive to anomalies, while a larger multiplier broadens the range





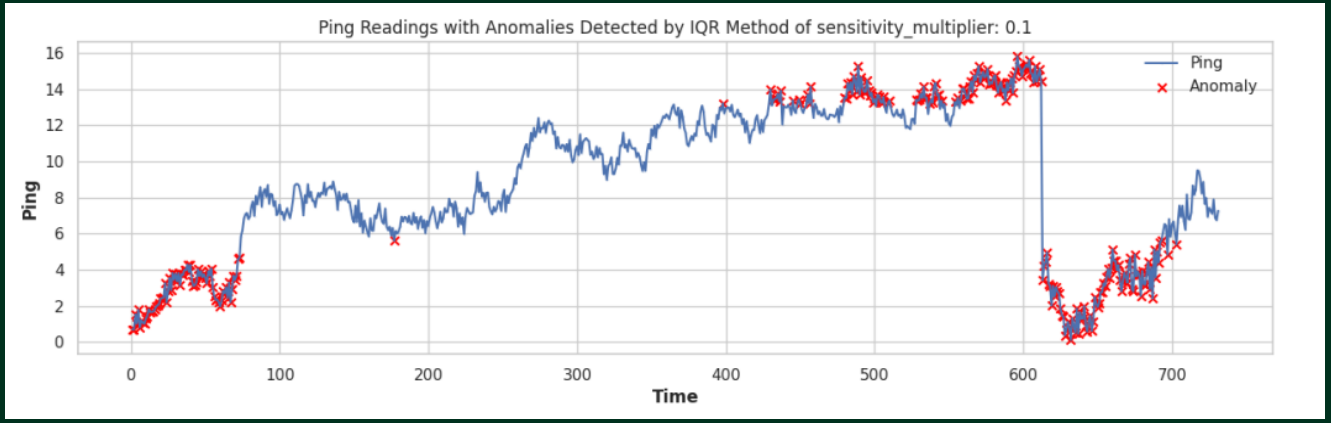
```
Q1 = df['ping'].quantile(0.25)
Q3 = df['ping'].quantile(0.75)
IQR = Q3 - Q1

sensitivity_multiplier = 0.7
lower_bound = Q1 - sensitivity_multiplier * IQR
upper_bound = Q3 + sensitivity_multiplier * IQR
df['anomaly_iqr'] = ((df['ping'] < lower_bound) | (df['ping'] > upper_bound))
```



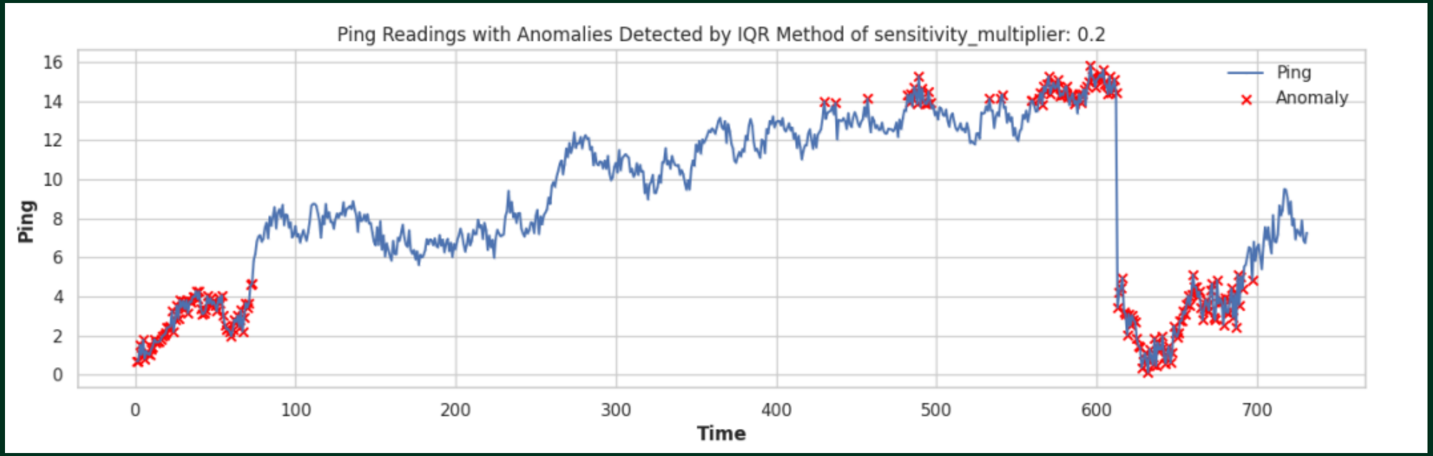
IQR Method : Anomaly Detection at sensitivity multiplier at 0.1

- Q1 and Q3 divide the data into four equal parts. Q1 is the value below which 25% of the data falls, and Q3 is the value below which 75% of the data falls.
- The range between Q1 and Q3, which contains the middle 50% of the data. It measures the spread of the central portion of the data.
- Adjusts the width of the anomaly detection bounds. A smaller multiplier results in a narrower range, making it more sensitive to anomalies, while a larger multiplier broadens the range



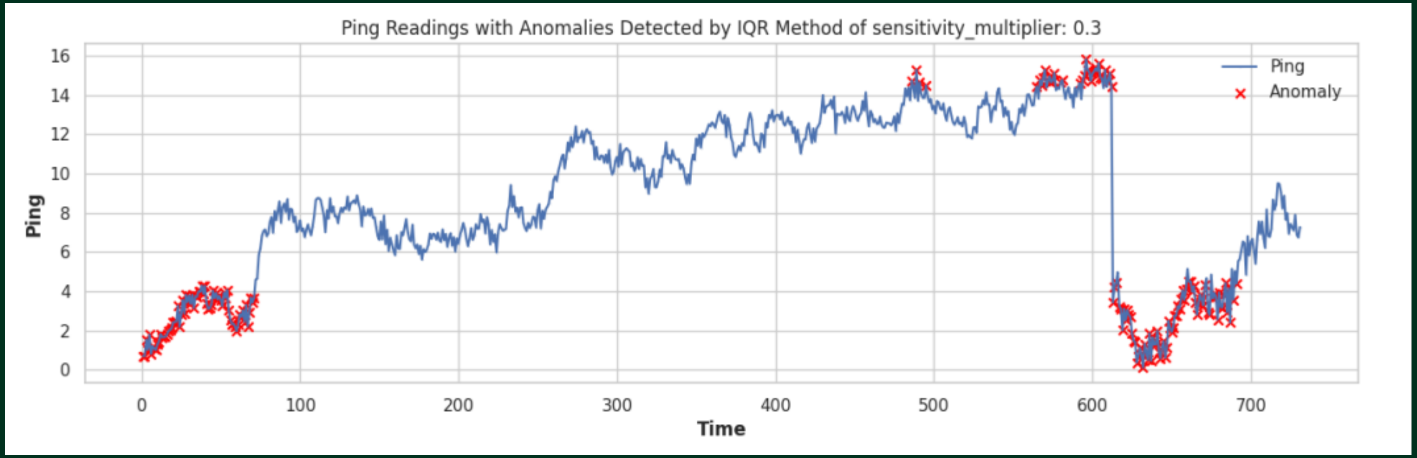
IQR Method : Anomaly Detection at sensitivity multiplier at 0.2

- Q1 and Q3 divide the data into four equal parts. Q1 is the value below which 25% of the data falls, and Q3 is the value below which 75% of the data falls.
- The range between Q1 and Q3, which contains the middle 50% of the data. It measures the spread of the central portion of the data.
- Adjusts the width of the anomaly detection bounds. A smaller multiplier results in a narrower range, making it more sensitive to anomalies, while a larger multiplier broadens the range



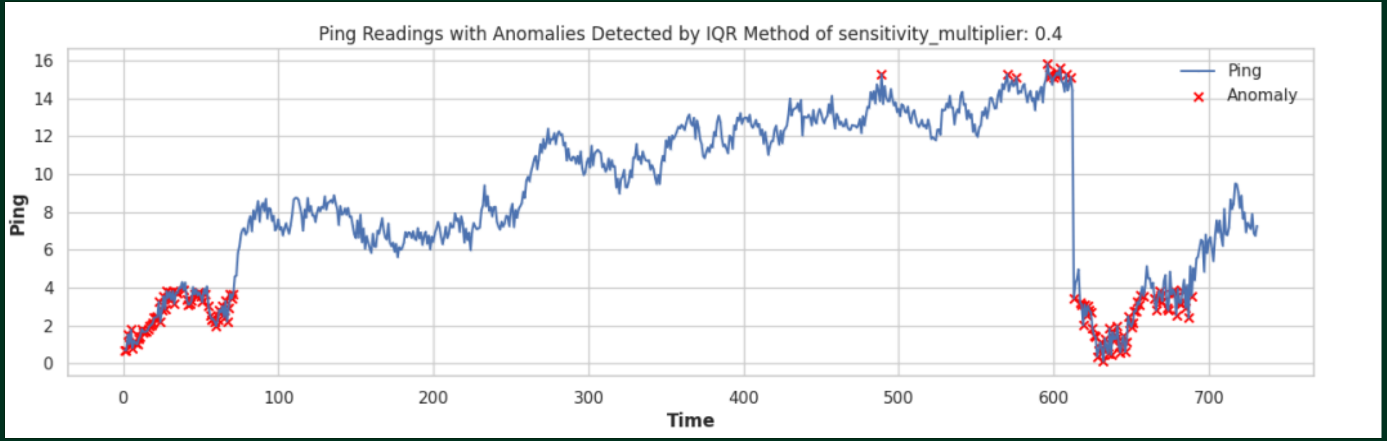
IQR Method : Anomaly Detection at sensitivity multiplier at 0.3

- Q1 and Q3 divide the data into four equal parts. Q1 is the value below which 25% of the data falls, and Q3 is the value below which 75% of the data falls.
- The range between Q1 and Q3, which contains the middle 50% of the data. It measures the spread of the central portion of the data.
- Adjusts the width of the anomaly detection bounds. A smaller multiplier results in a narrower range, making it more sensitive to anomalies, while a larger multiplier broadens the range



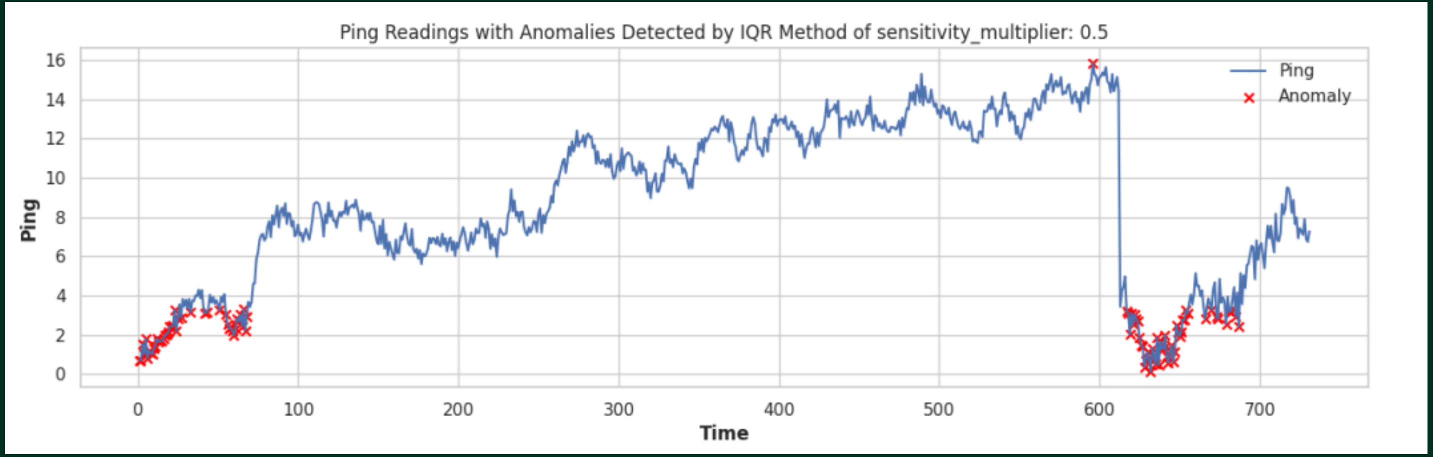
IQR Method : Anomaly Detection at sensitivity multiplier at 0.4

- Q1 and Q3 divide the data into four equal parts. Q1 is the value below which 25% of the data falls, and Q3 is the value below which 75% of the data falls.
- The range between Q1 and Q3, which contains the middle 50% of the data. It measures the spread of the central portion of the data.
- Adjusts the width of the anomaly detection bounds. A smaller multiplier results in a narrower range, making it more sensitive to anomalies, while a larger multiplier broadens the range



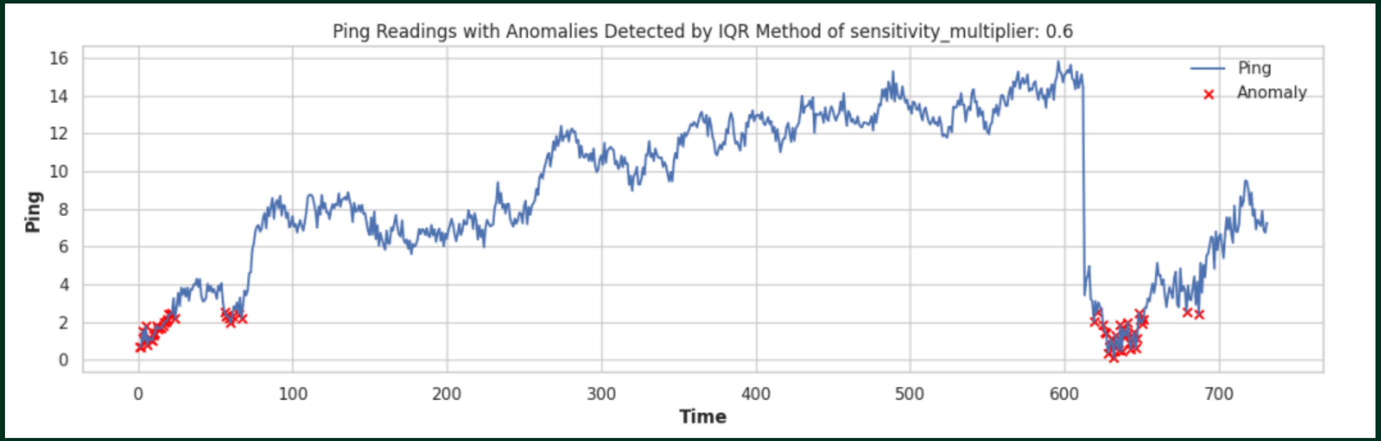
IQR Method : Anomaly Detection at sensitivity multiplier at 0.5

- Q1 and Q3 divide the data into four equal parts. Q1 is the value below which 25% of the data falls, and Q3 is the value below which 75% of the data falls.
- The range between Q1 and Q3, which contains the middle 50% of the data. It measures the spread of the central portion of the data.
- Adjusts the width of the anomaly detection bounds. A smaller multiplier results in a narrower range, making it more sensitive to anomalies, while a larger multiplier broadens the range



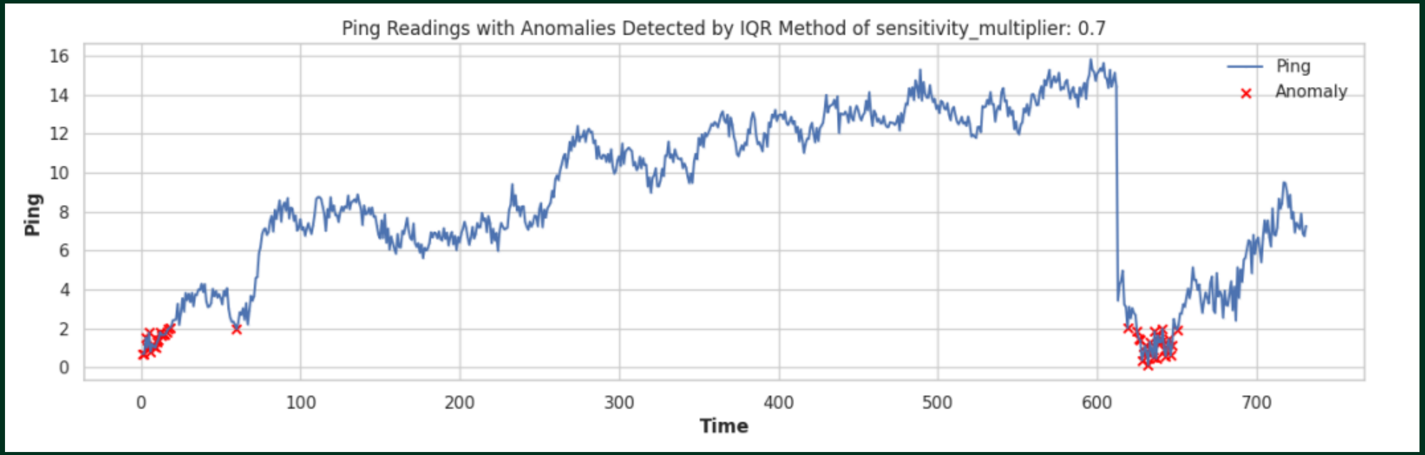
IQR Method : Anomaly Detection at sensitivity multiplier at 0.6

- Q1 and Q3 divide the data into four equal parts. Q1 is the value below which 25% of the data falls, and Q3 is the value below which 75% of the data falls.
- The range between Q1 and Q3, which contains the middle 50% of the data. It measures the spread of the central portion of the data.
- Adjusts the width of the anomaly detection bounds. A smaller multiplier results in a narrower range, making it more sensitive to anomalies, while a larger multiplier broadens the range



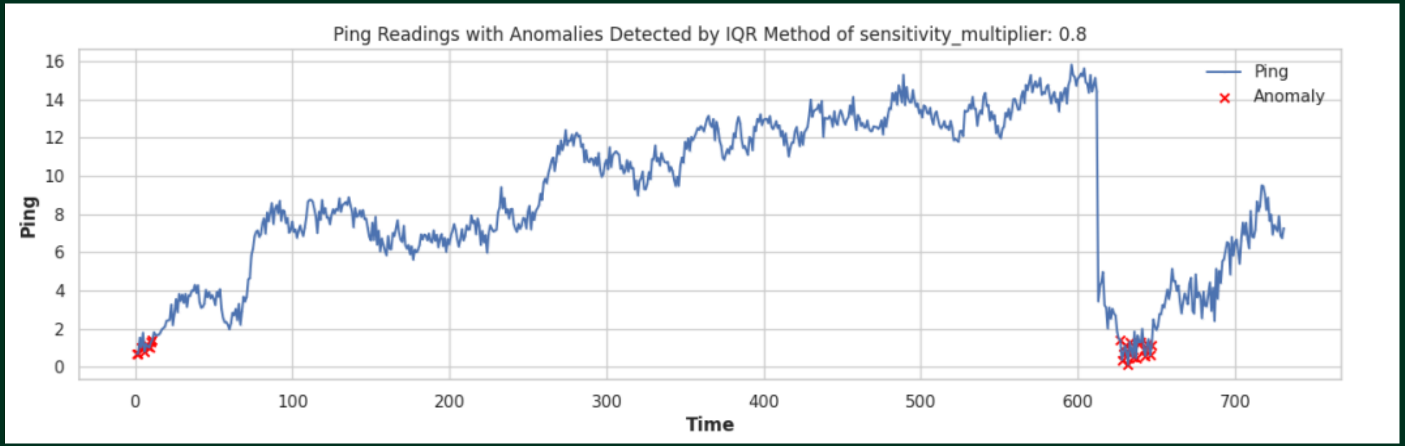
IQR Method : Anomaly Detection at sensitivity multiplier at 0.7

- Q1 and Q3 divide the data into four equal parts. Q1 is the value below which 25% of the data falls, and Q3 is the value below which 75% of the data falls.
- The range between Q1 and Q3, which contains the middle 50% of the data. It measures the spread of the central portion of the data.
- Adjusts the width of the anomaly detection bounds. A smaller multiplier results in a narrower range, making it more sensitive to anomalies, while a larger multiplier broadens the range



IQR Method : Anomaly Detection at sensitivity multiplier at 0.8

- Q1 and Q3 divide the data into four equal parts. Q1 is the value below which 25% of the data falls, and Q3 is the value below which 75% of the data falls.
- The range between Q1 and Q3, which contains the middle 50% of the data. It measures the spread of the central portion of the data.
- Adjusts the width of the anomaly detection bounds. A smaller multiplier results in a narrower range, making it more sensitive to anomalies, while a larger multiplier broadens the range



Conclusion

- In conclusion, we conducted exploratory data analysis on the given data and observed the data using histograms and boxplots. The data distribution lies between 6 to 8 and 12 to 14, and in the longer range, the major distribution is between 6 to 13. At the start, there are repetitive patterns with increasing ping values, which decrease to 2 at the 600 index.
- We also decomposed the series into trend, periodicity, and noise. We observed that the data contains periodicity but more than one cycle.
- For anomaly detection, we applied three different methodologies to detect anomalies in the data. Two of them are statistical: seasonal-trend decomposition using LOESS and the interquartile range, and one is machine learning-based: the isolation forest model. These are very strong methodologies for detecting anomalies in time series datasets.
- For evaluation, since actual outliers or ground truth were not available in the given datasets, we couldn't evaluate the predicted anomalies.

Future Plan

- In the future, we can apply more advanced methodologies from deep learning models, which require large amounts of data. For example, we can use a Siamese network, where two twin networks are used. We can train the model to identify patterns that are outliers and those that are not.
- We can also apply an Encoder-Decoder deep neural network. The Encoder's job is to transform the input into feature representations that can be easily distinguishable, and the Decoder's job is to regenerate the input. With this approach, we can train the Encoder-Decoder deep neural network so that the Decoder only knows how to generate normal signals. If the error is very high, it means the given input series contains outliers.

EVERGREEN

INTRODUCING INSIGHT GLOBAL’S MANAGED SERVICES CAPABILITIES