Aalto University
School of Electrical
Engineering

ELEC-E7130 Internet Traffic Measurements and Analysis

Assignment 5. Data Analysis

Name: Xingji Chen

Student ID: 101659554

E-mail: xingji.chen@aalto.fi

Aalto University
School of Electrical
Engineering

Task 1: Understanding different plots

1. Autocorrelation plot
   - Y-axis: Autocorrelation coefficient (ranging from -1 to 1).
   - X-axis: Time lag between data points.
   - Usage: Autocorrelation plots are used to assess the similarity between a signal and its own delayed version. They are particularly useful in time series analysis for detecting patterns, seasonality, and dependencies in the data.
   - Limitations: Autocorrelation plots may not provide a complete understanding of complex data relationships. They assume that the data is smooth, which does not always hold true in real data.

2. Boxplot
   - Y-axis: Data values.
   - X-axis: Category or Group.
   - Usage: Boxplots are used to visualize the distribution of a data set, showing medians and possible outliers. They are used to identify spread and skewness in data, as well as to compare multiple groups.
   - Limitations: Boxplots do not provide a detailed view of the data distribution, making them less specific in showing the exact shape of the data. They may not be applicable to small data sets.

3. Lag plot
   - Y-axis: The data value at the moment t+1.
   - X-axis: Data values at moment t.
   - Usage: Lag plots are used to assess the relationship between a data point at a given moment and the same data point at a previous moment. They help to detect serial correlations or patterns in time series data.
   - Limitations: Hysteresis plots are most effective for detecting simple linear dependencies, but may not capture more complex relationships or nonlinear patterns.

4. Parallel plot
   - Y-axis: Data values or dimensions.
   - X-axis: Parallel axes representing different characteristics or attributes.

- Usage: Parallel coordinate plots are mainly used for multivariate data visualization. They show individual data points or clusters with different differences along multiple variables and are suitable for clustering and outlier detection.
- Limitations: As the number of dimensions increases, parallel coordinate plots become less specific and interpretable. They may not be effective in capturing relationships between multiple variables or in providing detailed distributional information.

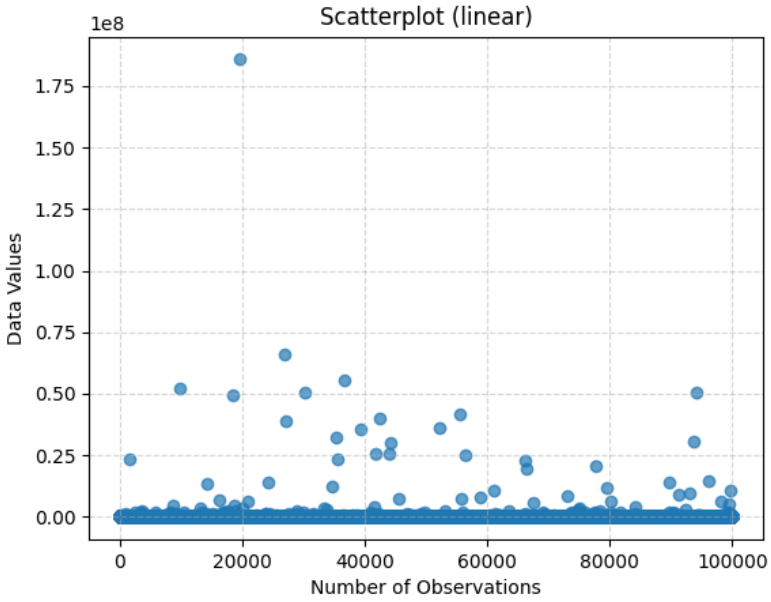
5. Scatter Matrix Plot
   - Y-axis: A variable.
   - X-axis: Another variable.
   - Usage: Scatter matrix plots display paired scatter plots of multiple variables. They are used to visualize relationships between variables in a data set, identify correlations and assess data distribution.
   - Limitations: As the number of variables increases, scatter matrix plots can become confusing and less interpretable. They do not provide a complete picture of high-dimensional data and may not reveal nonlinear relationships.

Aalto University
School of Electrical
Engineering

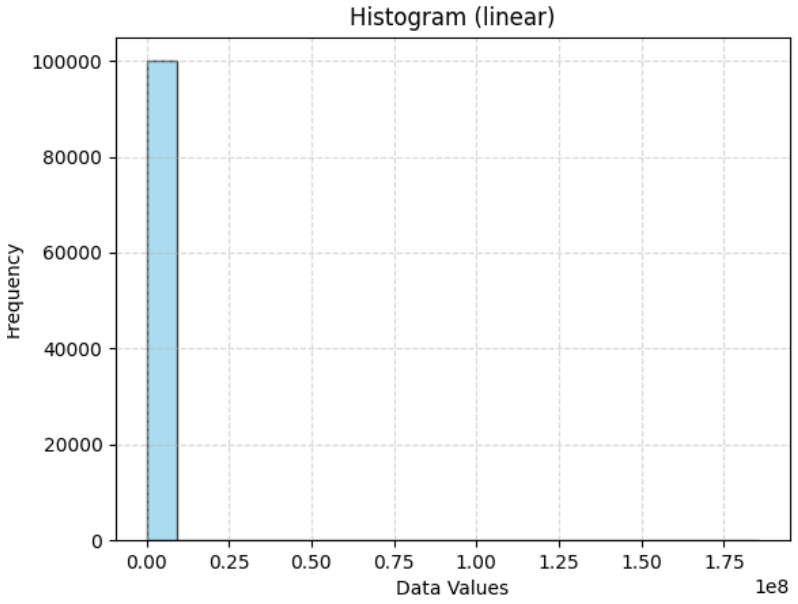
Task 2: Plot data

1. Plot the flow data

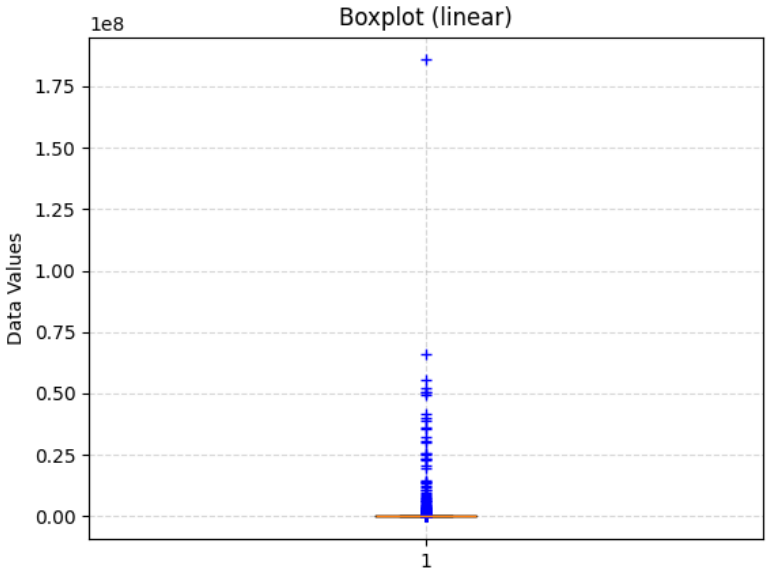
- Scatterplot (Number of observations will reside on X-axis)



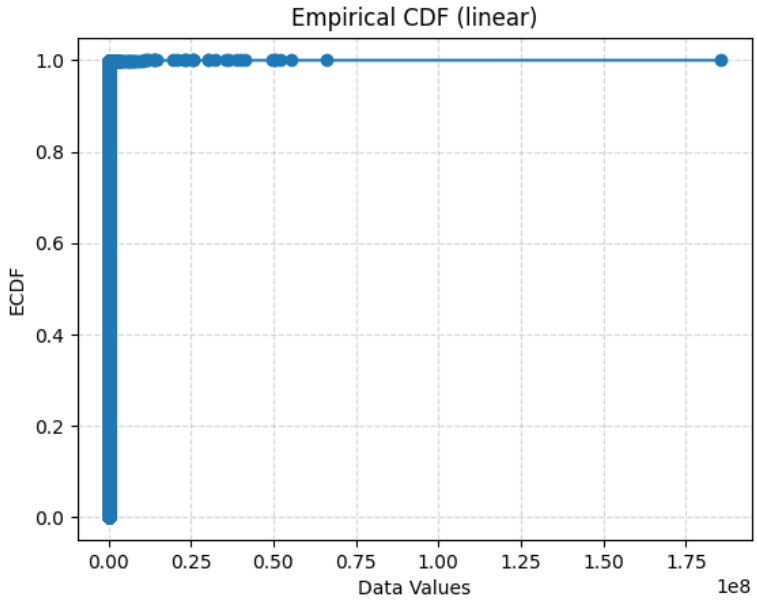
- Histogram (Using a suitable number of bins)



- Boxplot



- Empirical CDF of the variable



- Scatterplot:

  The distribution of the data is mainly concentrated in the lower range of data values. However, there are some data points, especially those with very high data values, that show significant outliers.

- Histogram:
  The vast majority of data values are clustered in the lowest range, meaning that most of the flow data is relatively small. Only a small number of data points have high data values, which is consistent with the outliers observed in the scatterplot.

- Boxplot:
  The median and quartiles of the data are in the very low range, indicating that most of the data is very concentrated. The presence of many outliers, which are in the high range of data values, further confirms the presence of a few very high data values.

- Empirical CDF:
  About 80% of the data is concentrated in the lowest data value range, which is consistent with the histogram observations. The growth of the data slows down to near flatness as the data values increase, implying that the frequency of high data values is very low.

2. Describe the distributions choosing variables.

- Maximum values
  The maximum value can be found quite intuitively in the scatterplot, the highest point is the maximum value of the set of data, which is approximately 1.82*10^8. The maximum value provides us with an upper limit to the data stream. Knowing this, we can better assess where the other data points are relative to the maximum value. In network traffic, the maximum value can help us understand the peak demand of the system so that we can allocate and scale resources appropriately. In some applications, the maximum value may imply some sort of overload or anomaly. Knowing the maximum value can help us detect and respond to possible problems earlier.

- Mode
  The mode represents the value that appears most frequently in the data, thus it can quickly provide us with a notion of the most common state or value of the flow. If a particular flow value appears as the mode very frequently, it might suggest that the flow remains stable for the majority of the time. Conversely, if there is no distinct mode in the data or there are multiple modes, it might indicate that the flow changes frequently. Knowing the mode of the flow data can help organizations or system administrators allocate resources more effectively. For example, if the flow value indicated by the mode is the most

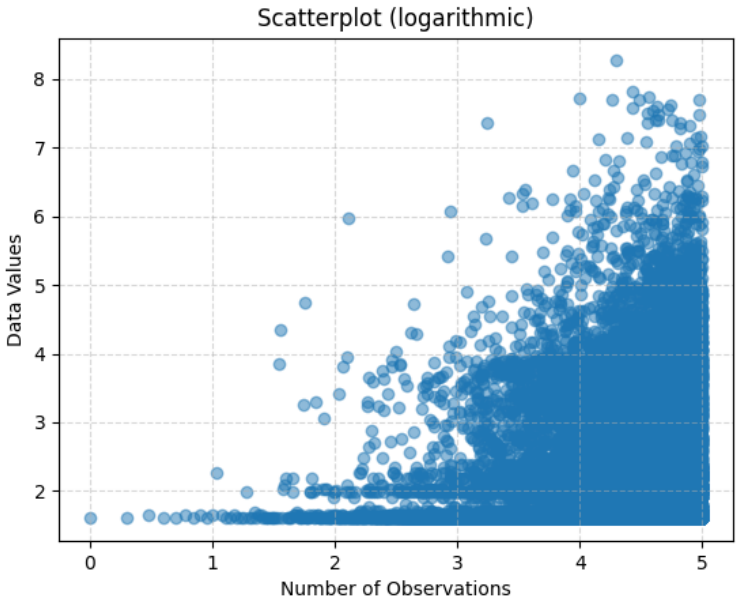
common, it ensures that the system operates most efficiently under that flow. A specific mode might be associated with certain patterns or behaviors. For instance, if a website's traffic mode is at a particular value, it might imply that this is the standard number of visitors, and any deviation from this pattern might warrant further analysis.

- Median

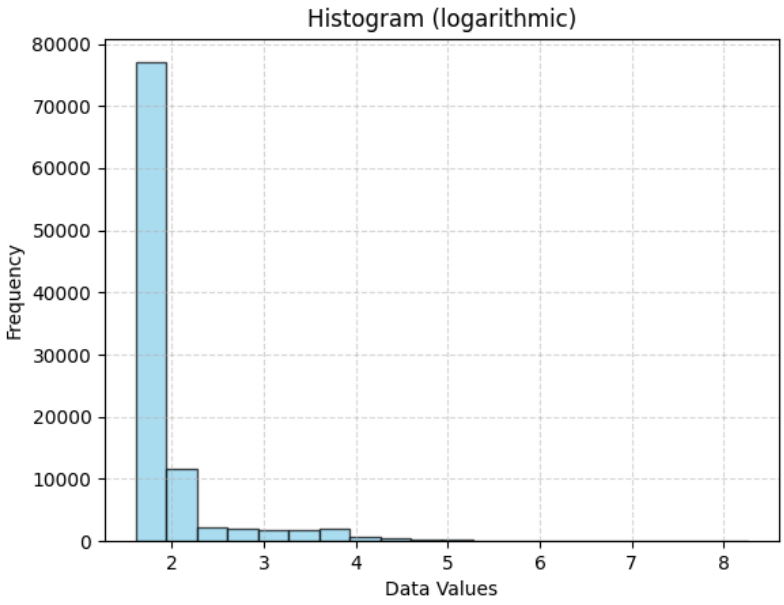
  The median represents the central tendency of data and can be used to determine the typical value of flow data. Unlike the mean, the median is not influenced by outliers, thereby offering a more stable and accurate representation of the data's center. In flow data, there might be sudden events causing a short-term surge or sharp decrease in flow. The median, unlike the mean, is not heavily affected by these outliers, thus better reflecting the general trend of the data. The median divides the dataset into two equal parts, where half of the data values are below the median and the other half are above. This aids in understanding the distribution of the data. The median can assist decision-makers in identifying and interpreting the typical behavior of the flow, leading to better planning and decision-making.

3. Replot data using logarithmic values

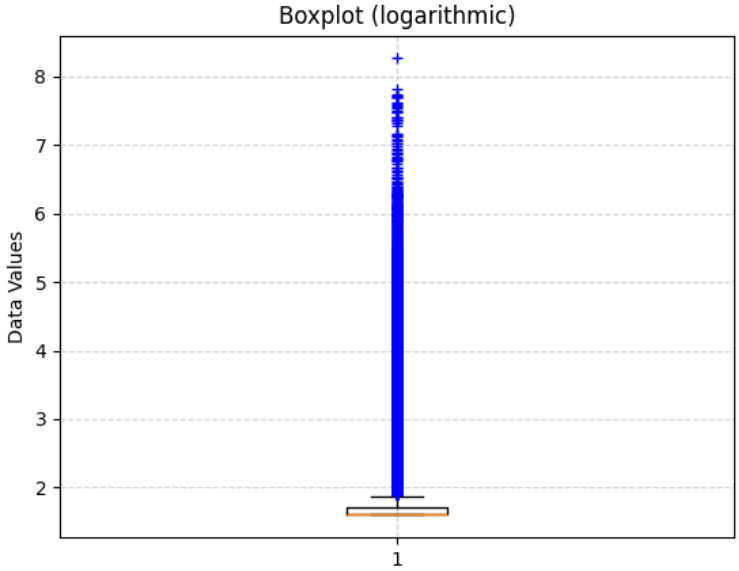
- Scatterplot (Number of observations will reside on X-axis)



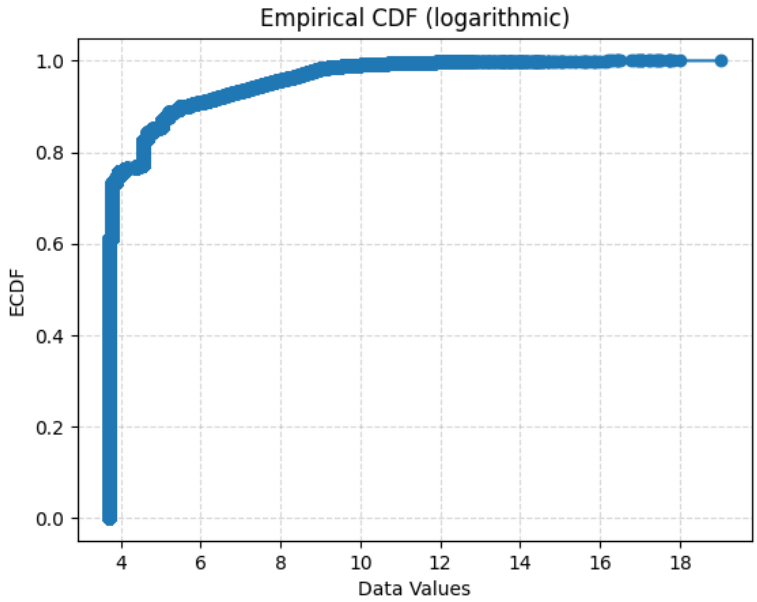
- Histogram (Using a suitable number of bins)



- Boxplot



- Empirical CDF of the variable



- Scatterplot:

  The data shows a growing trend from the bottom left to the top right, which means that the data values are increasing with time. When the number of observations is small, the data values are sparsely distributed; while when the number of observations" increases, the data values become denser.

- Histogram:

  Most of the data values are clustered in the range of 2-3, which is the most common range of data values. The frequency decreases rapidly as the data values increase, indicating that higher data values are less likely to occur.

- Boxplot:

  The median of the data is in the lower range. However, there are a large number of outliers that are well above the upper quartile of the data.

- Empirical CDF:

  Approximately before a data value of 8, the cumulative distribution increases rapidly, with over 80% of the data points below this value. After a data value of 8, the growth trend slows down and approaches a plateau.

- Why and when it is more suitable to use the logarithmic values?
  When some data sets have a wide range of values, from very small to very large, this can cause problems in statistical analysis. That is, when the data have significant heteroskedasticity, logarithmic transformation can stabilize the variance. At the same time, the logarithmic scale can represent a wide range of values in a compact manner. Especially when looking at data that spans several orders of magnitude, logarithmic scales can provide more intuitive results.

4. Conclusion

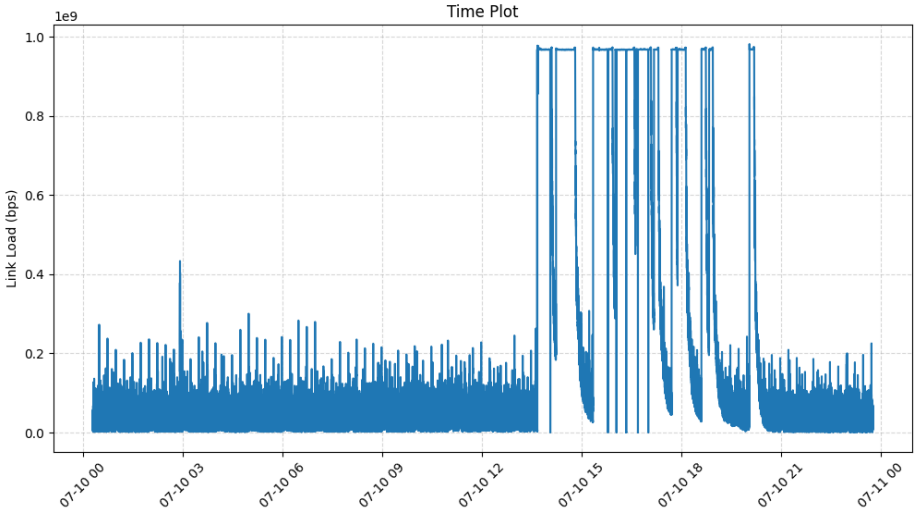
Each method provides a different perspective for describing and understanding the data. Therefore, choosing the best method depends on what kind of information we want to get from the data. If the goal is to understand the overall distribution and concentration trends of the data, histograms and box plots are very appropriate. They can quickly tell us where the bulk of the data is concentrated and whether there are any outliers. However, if we want to understand the relationship between two variables in detail, then scatter plots are more appropriate. ECDF plot shows the cumulative distribution of the data values, and we can confirm the range in which most of the data lies.

Aalto University
School of Electrical
Engineering

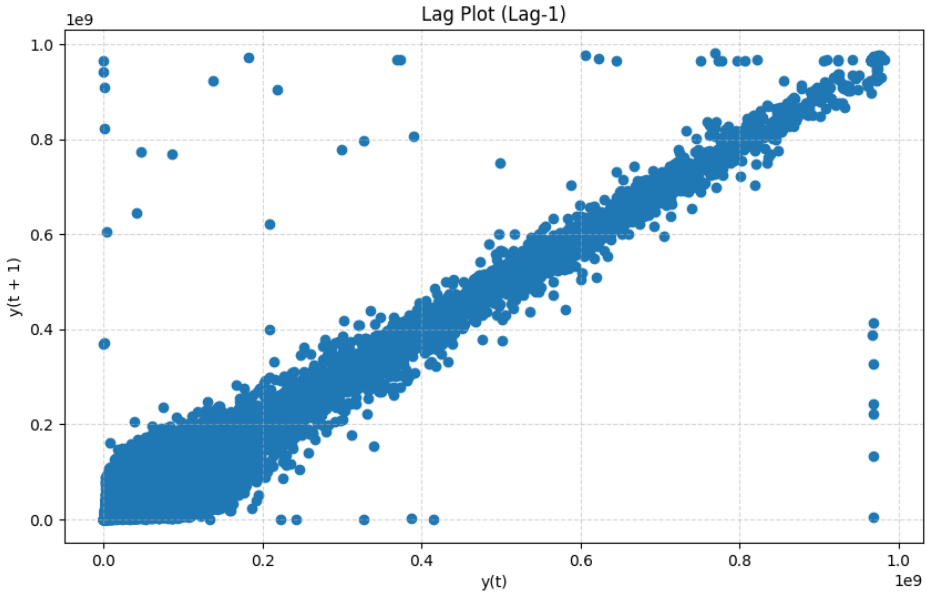
Task 3: Link loads

1. Linkload-1

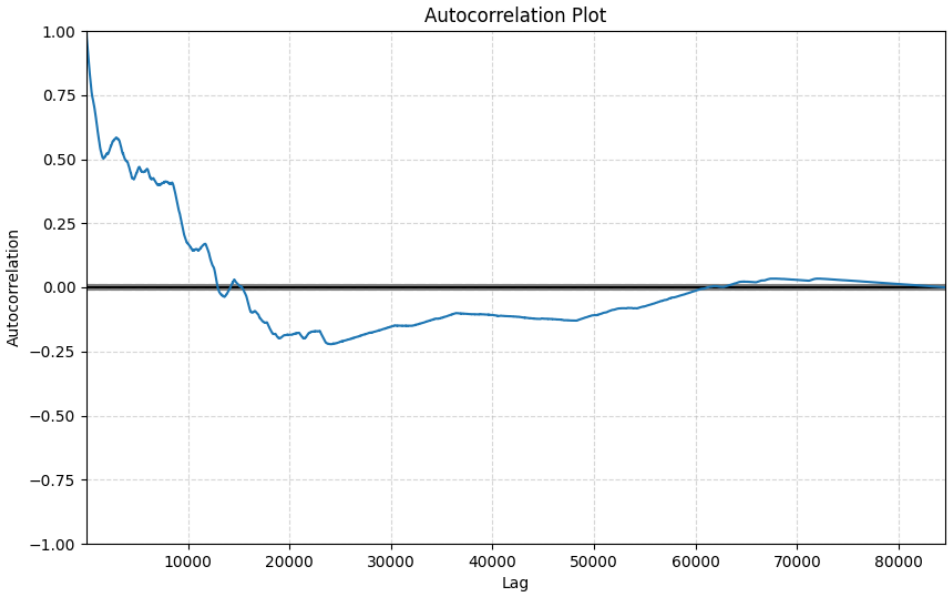
- Time plot



- Lag plot (lag-1)

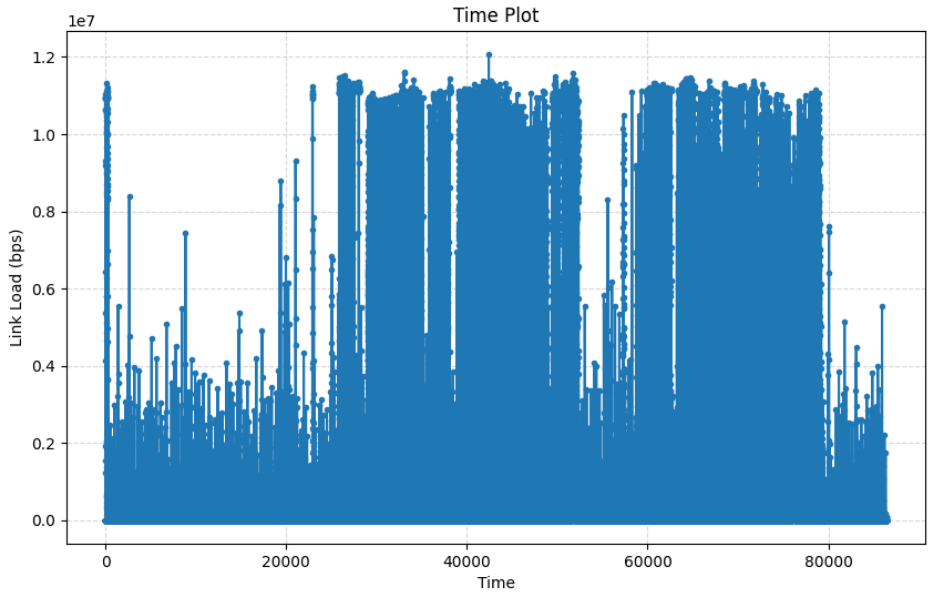


- Correlogram (i.e. autocorrelation plot)

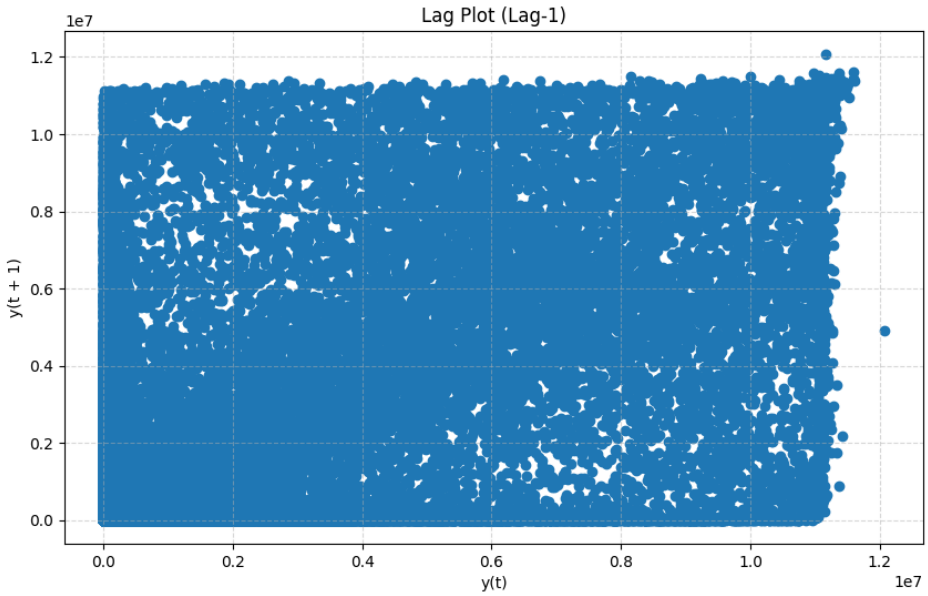


2. Linkload-2

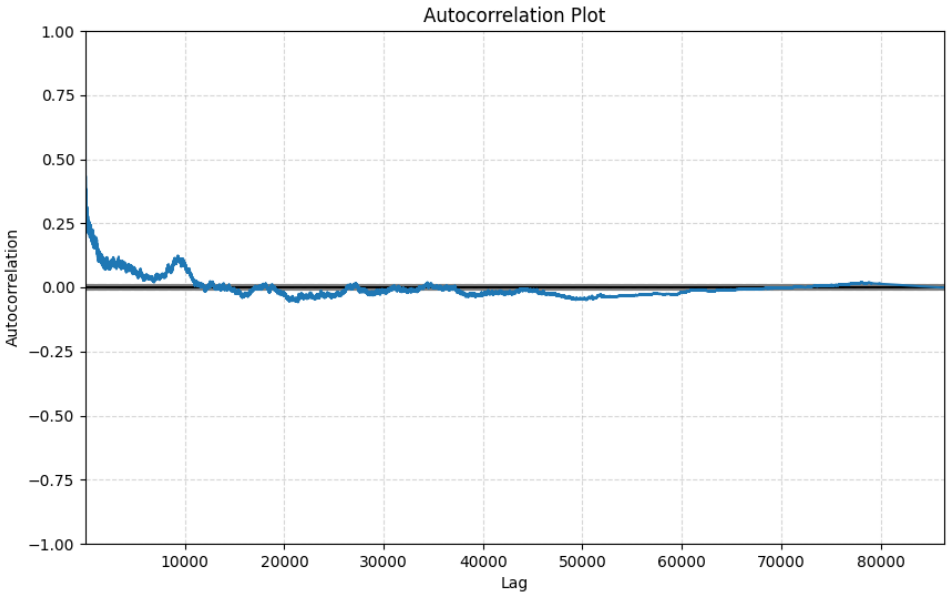
- Time plot



- Lag plot (lag-1)

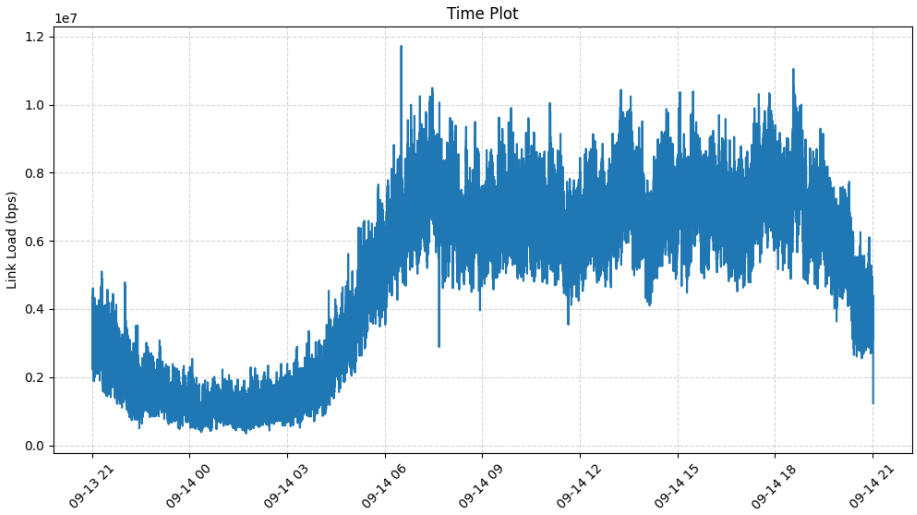


- Correlogram (i.e. autocorrelation plot)

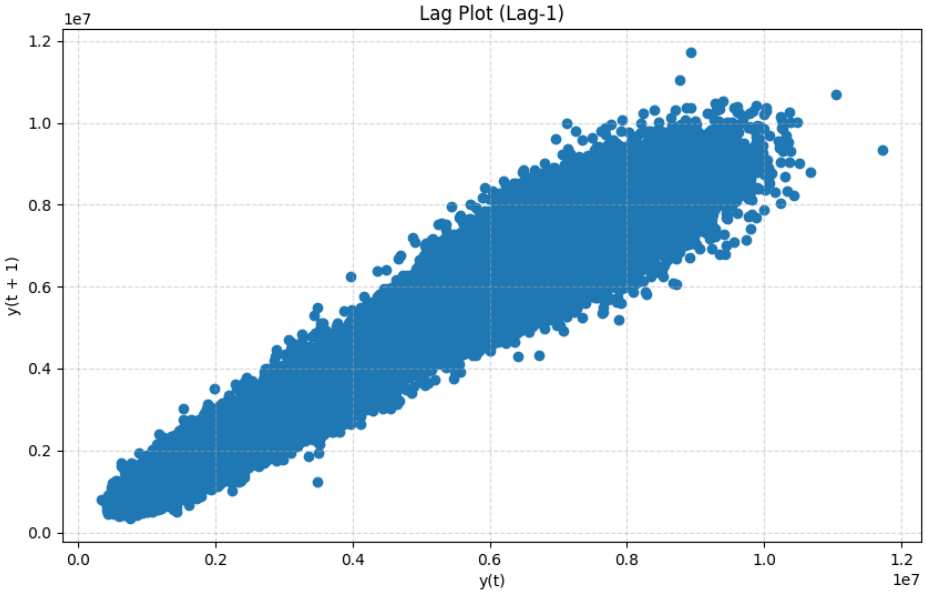


3. Linkload-3

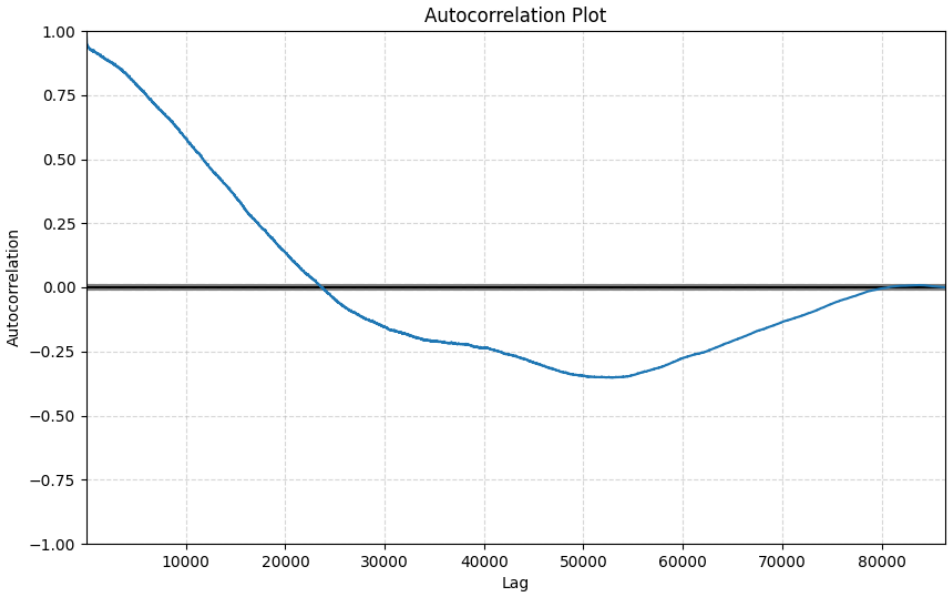
- Time plot



- Lag plot (lag-1)

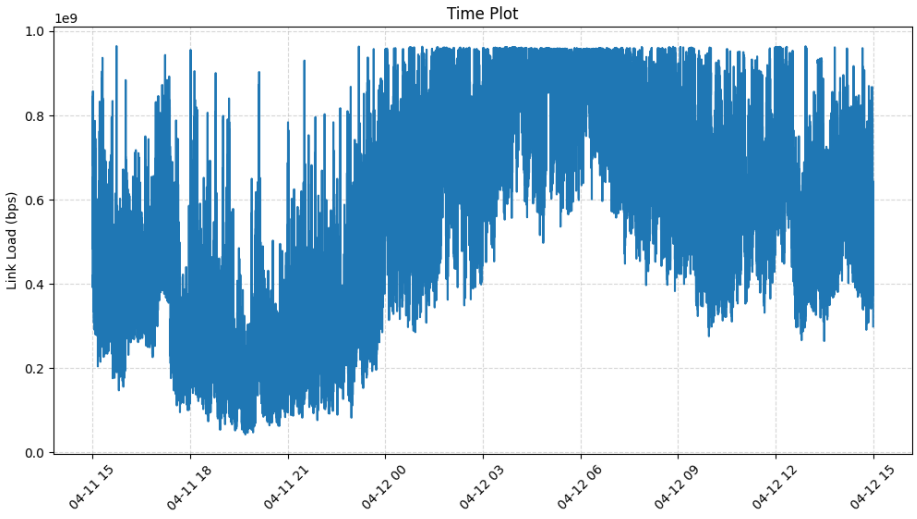


- Correlogram (i.e. autocorrelation plot)

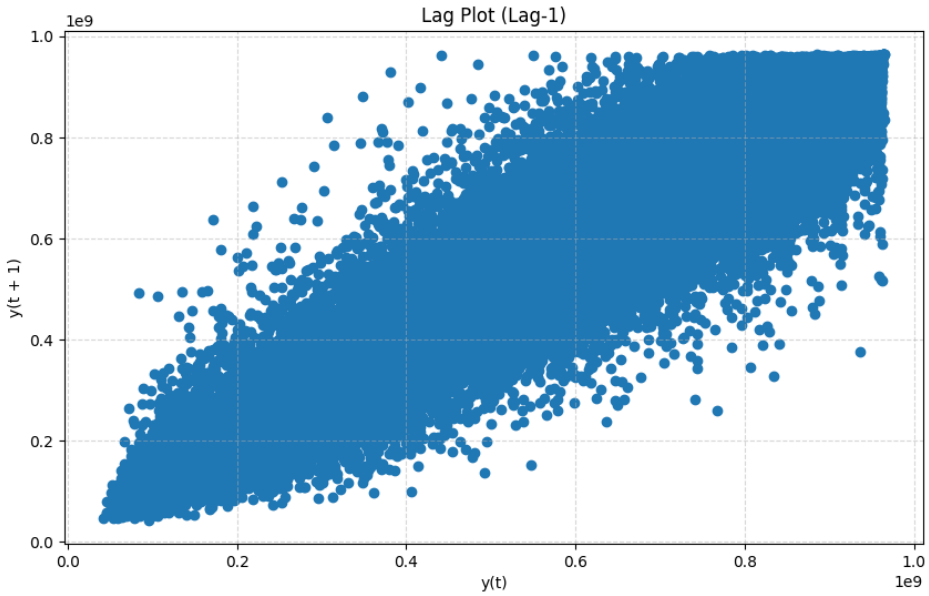


4. Linkload-4

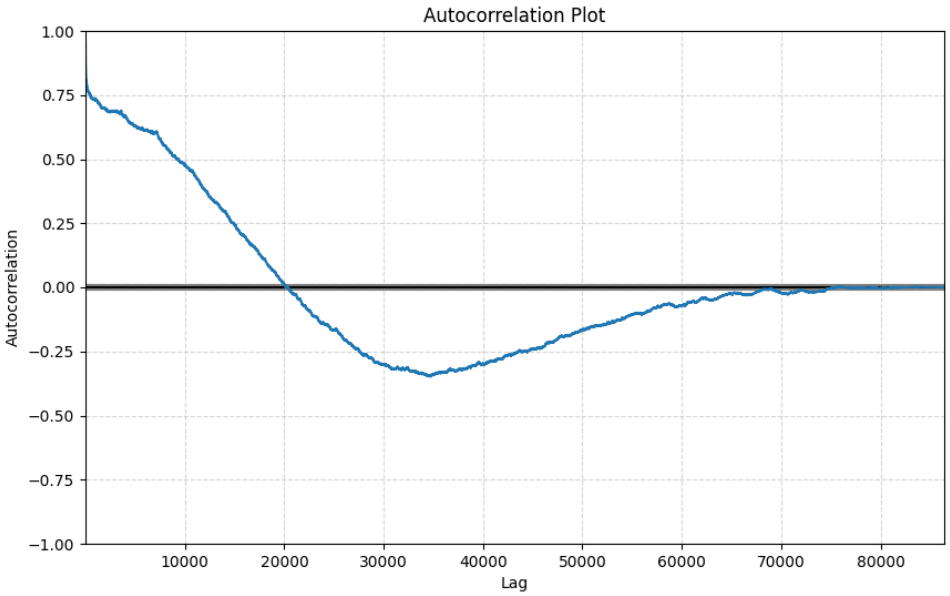
- Time plot



- Lag plot (lag-1)



- Correlogram (i.e. autocorrelation plot)



5. Inspect the data results

1) Linkload-1

- As can be observed from time plot, the mean and variance of the data are relatively stable most of the time, despite the presence of peaks. This is an indicator of time series stability.
- From lag plot (lag-1), we can see a strong correlation in the data, which means that previous values (specifically those from a previous point in time) are strongly correlated with the current values. This can be thought of as short-term memory, where one or several previous observations help to predict the current value.
- Autocorrelation plot gives us a more in-depth view. At initial lag values, there is a high degree of positive autocorrelation, which further confirms the short-term memory observation. However, this correlation gradually decreases as lag increases. This pattern may be indicative of long-term memory.

2) Linkload-2

- The data is not completely stable as we can observe clear fluctuations and periodic structure. The variance of the data increases significantly during certain time periods.
- Lag plot shows a dispersed pattern with no clear trend or shape, which implies that the data may be random and that the short-term memory of the data is weak, i.e., there is no strong relationship between previous values and their subsequent values.
- As lag increases, the autocorrelation decays and approaches zero after a certain point. This indicates a weak long-term memory effect on the data.

3) Linkload-3

- The time plot shows a clear non-linear trend in the data, showing a pattern of increasing and then decreasing. The data shows a high degree of volatility within its range.
- Lag plot shows a clear positive correlation. When the value at moment t-1 is lower, the value at moment t is also lower; and vice versa. This positive correlation indicates the presence of short-term memory.
- In the autocorrelation plot, we can see that in the initial lag, the autocorrelation decreases rapidly. This is further evidence that the data has a short-term correlation or memory. Subsequently, the autocorrelation fluctuates around zero in the longer lag, indicating that the long-term memory is not strong.

4) Linkload-4
   - The time plots show that the data has significant seasonality and volatility.
   - The Lag plot demonstrates a broad distribution of data points with no clear linear pattern. This means that there is no direct linear relationship between the value at a single time point and its value at the previous time point.
   - The autocorrelation plot shows that there is significant autocorrelation in the initial lag, which suggests that the data has short-term memory. As the lag increases, the autocorrelation decreases, but there is still some degree of autocorrelation. This may indicate that the data also has some long-term memorability.

6. Understanding of each data set

1) Linkload-1
   - The Time Plot showed an increasing trend, followed by a notable peak in the middle, and then a stabilization of the trend.
   - The Lag Plot displayed a more dispersed distribution of points, but with higher density in some regions, indicating some level of memory in the data.
   - The Autocorrelation Plot indicated strong autocorrelation in the initial lags, which dropped off rapidly afterward.

2) Linkload-2
   - The Time Plot had data starting from a low point and gradually increasing, then maintaining at a relatively higher steady state.
   - The Lag Plot showed a more concentrated distribution of points, suggesting that previous values could potentially be helpful in predicting subsequent ones.
   - The Autocorrelation Plot depicted a diminishing autocorrelation in the initial lags but still maintained some degree of autocorrelation.

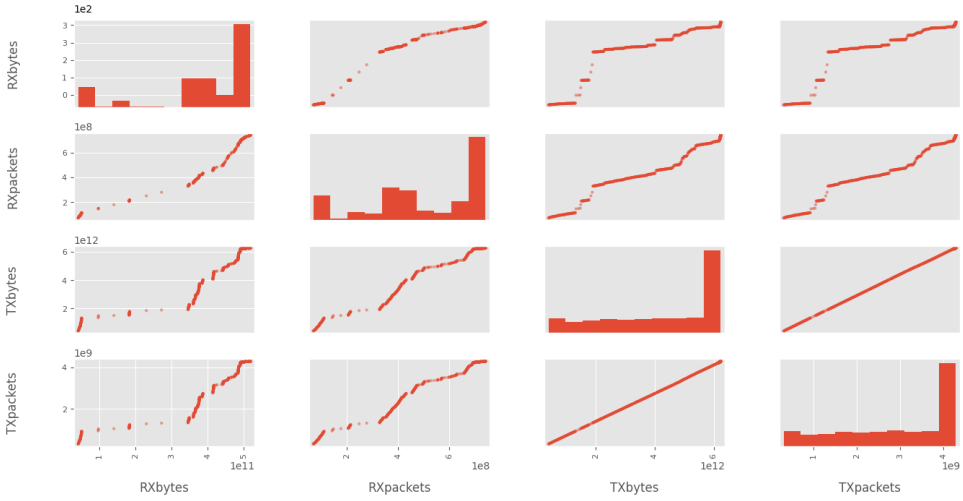
3) Linkload-3
   - The Time Plot presented a more volatile data pattern and it has clear ascents and descents.
   - The Lag Plot had a widespread distribution of points without a clear linear pattern.
   - The Autocorrelation Plot showed short-term and long-term memory in the data, with significant autocorrelation in the initial lags that dwindled over time.

4) Linkload-4
   - The Time Plot showcased stability in the data during certain periods, interspersed with noticeable rises and falls.
   - The Lag Plot exhibited a broad distribution of points, without a clear linear relationship.
   - The Autocorrelation Plot revealed short-term memory in the data, and as the lags increased, there still remained some autocorrelation, suggesting some form of long-term memory in the data.

Aalto University
School of Electrical
Engineering

Task 4: Pairs plot

1. Plot the pairs plot for such values.



2. Answer the following questions:

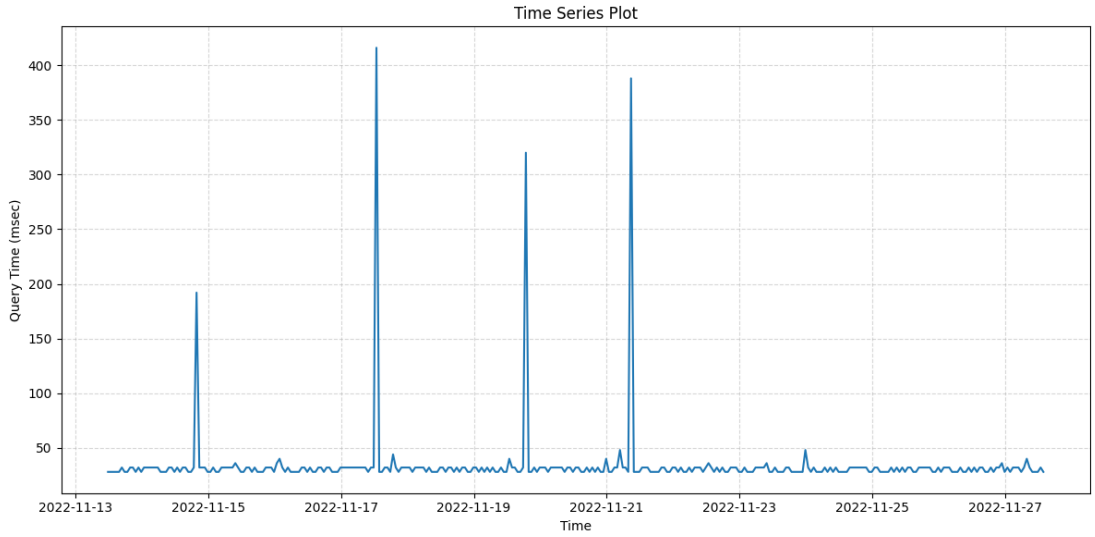
   1) Which variables correlate most to each other?
      - From the pairs plot, the relationship between TXbytes (transmitted bytes) and TXpackets (transmitted packets) is very linear, indicating that they are highly correlated. Similarly, RXbytes (received bytes) and RXpackets (received packets) also have a similarly high linear relationship, suggesting that they are also highly correlated.

   2) Let's assume that you decide to remove one particular column to reduce the computation load of data handling. Based on the pairs plot, what would the column be, and why?
      - If I were considering reducing the computational load of data processing and chose to delete a column, I would consider deleting either TXpackets or RXpackets.This is because, as mentioned earlier, there is a very high correlation between TXbytes and TXpackets, and between RXbytes and RXpackets. Not much information is lost by deleting one of them, as they have a strong correlation with the other column and can be deduced from the other column.

Aalto University
School of Electrical
Engineering

Task 5: Understanding time series concepts

1. Plot the time series.



2. By observing and analyzing the plot, answer the following questions:

   1) Is there any trend or seasonality?
      - From the time series plot, there is no clear long-term trend in the data, but there are a few distinct peaks. These peaks may indicate some sort of seasonal or recurring pattern, but the short time frame provided makes it difficult to determine if there is a fixed seasonal pattern.

   2) Is the time series stationary?
      - Time series are not exactly stationary. A stationary time series means that its statistical properties are constant in time. On this plot, there are several significant peaks, although most of the time query times remain in the lower range. This means that its mean and variance may vary at different times and therefore the time series is not completely stationary.