1) Power Consumption (in GW) over Time:

We can see that there's been a rapid increase in power consumption and annaualised power consumption over time. They remained relatively stable in the initial years and then witnessed a significant spike. These fluctuations might be attributed to various factors such as changes in Bitcoin's price, mining profitability, advances in mining hardware, or the overall growth of the Bitcoin network.

2) Annualized power consumption (in TWh) over time:  
  
Similar to the power consumption plot, the annualized consumption showcases a consistent growth pattern, with a significant increase in the recent years. There's a noticeable surge in the recent years, which might be attributed to the increasing popularity and acceptance of Bitcoin, leading to more miners joining the network and thus consuming more power.

3) Estimated efficiency J/Th:

This plot showcases a general decreasing trend over time, which means that mining operations are becoming more energy-efficient meaning they require fewer joules of energy to produce a terahash. The steady decrease can be attributed to technological advancements in mining hardware.

There are periods where the efficiency remains relatively constant, possibly indicating times when there were no significant advancements in mining technology.

4) Carbon emissions (in MtCO2e) over time.

Next, let's analyze the environmental impact of Bitcoin mining by visualizing the carbon emissions (in MtCO2e) over time.

Emission Ranges:

Hydro-only Emissions: When considering hydroelectric power as the sole source, the emissions are significantly lower. This emphasizes the cleaner nature of hydroelectric power.

Coal-only Emissions: On the other end, when considering coal as the sole power source, emissions are much higher. This highlights the environmental concerns associated with coal-based power sources.

Estimated Emissions: The estimated emissions lie between the hydro-only and coal-only values, suggesting a mix of power sources.

All three emission metrics showcase a steady increase over time, aligning with the growth in power consumption and annualized consumption trends observed earlier. Switching to cleaner energy sources can drastically reduce the carbon footprint of the Bitcoin network.

Next, let's visualize the emission intensity over time. This will show us how efficiently Bitcoin mining has been conducted in terms of carbon emissions relative to power consumption.

7)Emission Intensity over Time:

The emission intensity remained relatively stable during the initial period. As time progressed, we can observe fluctuations in the emission intensity. This suggests variability in the efficiency of Bitcoin mining in terms of carbon emissions relative to power consumption. There’s a rising trend in emission intensity. This could be attributed to increased mining activities, possibly relying more on non-renewable energy sources.

Finally, let's visualize the hash rate over time. The hash rate indicates the computational power used in mining and transaction confirmations. It will reflect the growth in mining activity and the evolution of mining technology.

8)Hash Rate (in MH/s) over Time:

The hash rate has exhibited a massive growth over the years. This indicates an exponential increase in the computational power dedicated to Bitcoin mining. The growth in the hash rate can be attributed to the evolution of mining technology, from basic CPU mining in the early days to GPU, FPGA, and ASIC mining in more recent times.

A higher hash rate means the Bitcoin network is more secure against attacks. However, it also suggests higher energy consumption, as observed in the power consumption plots.

Overall, the data and plots show the evolution of Bitcoin mining over time, from its nascent stages to its current state. The increase in power consumption, annualized consumption, carbon emissions, and hash rate points to the growing popularity and scale of Bitcoin. However, the environmental implications, especially in terms of carbon emissions, highlight the importance of using sustainable and renewable energy sources for mining activities.

--------------------------

What is Stationarity?

In the context of time series analysis, a stationary process (or a stationary series) is one whose statistical properties (like mean, variance, and Standard deviation) are all constant over time. In simpler terms, the series doesn't have any predictable patterns in the long-term. Fluctuations in the series over time are purely random.

There are two main reasons why stationarity is important:

1. Predictability: Many forecasting methods assume or require the time series to be stationary. If a time series has a particular behavior over time, there is a very high probability that it will follow the same in the future. As a result, it becomes easier to predict.

2. Model Simplicity: Stationary processes are easier to model. Non-stationary data often require additional measures (like differencing or transformations) to make them stationary, which can complicate the modeling process.

Why Choose Stationarity Analysis?

1. Model Assumptions: Many time series models, like ARIMA (AutoRegressive Integrated Moving Average), assume that the underlying data is stationary. If the data isn't stationary, the predictions made by these models won't be reliable.

2. Detecting Root Causes: Stationarity analysis can help in identifying underlying factors causing a change in the time series. For instance, a sudden change in variance might indicate an external factor (like a market crash or a major event) affecting the time series.

3. Improving Model Performance: Making a time series stationary can lead to better model performance as the series becomes more predictable.

Typically, stationarity analysis plots may include:

1. Original Time Series Plot: This shows the raw data over time.

2. Rolling Mean: This is the moving average of the time series. If it remains constant over time, it's an indication that the series might be stationary in terms of mean.

3. Rolling Standard Deviation: This is the moving standard deviation of the time series. A constant rolling standard deviation suggests that the series might be stationary in terms of variance.

**Histograms**

annualised consumption (in TWh):

This show a right-skewed distribution, indicating that most of the values are clustered around the lower end with a few higher values. This suggests that for a majority of the time, the annualized consumption values were relatively low with occasional spikes.

power ( in GW):

Similar to the annualized consumption, this histograms are right-skewed, indicating most of the power consumption values are on the lower end with occasional higher values.

Efficiency (Lower bound, Estimated, Upper bound in J/Th):

This histogram is right-skewed, suggesting that for a majority of the time, the efficiency values were low, with occasional periods of higher efficiency.

Carbon Emissions (Hydro-only, Estimated, Coal-only in MtCO2e):

The histograms for these features are also right-skewed. This indicates that for a significant duration, the carbon emissions were on the lower side, with occasional periods of higher emissions.

Emission Intensity (gCO2e/kWh):

This histogram is more bell-shaped, indicating a somewhat normal distribution. Most of the values are clustered around the middle, suggesting a consistent emission intensity for a significant duration.

Hash rate (MH/s):

The histogram is right-skewed, indicating that for a majority of the time, the hash rate was relatively low, with occasional spikes representing periods of increased mining activity.

From these histograms, we can infer that many of the features have right-skewed distributions, indicating that higher values are less frequent and occur as outliers or spikes. These spikes could be attributed to specific events, technological advancements, or market dynamics related to Bitcoin.

Let's take a statistical approach to detect these outliers.

One common method for outlier detection is the IQR method. In this method:

* Calculate the Interquartile Range (IQR) as the difference between the third quartile (Q3) and the first quartile (Q1).
* Define the lower boundary as Q1−1.5×IQR and the upper boundary as Q3+1.5×IQR.
* Any data point outside these boundaries is considered an outlier.

Capping is a method used to limit extreme values in the statistical data to reduce the effect of possibly spurious outliers. By doing this, we can make the data more robust and reduce the influence of outliers on our analysis.

The capping method involves:

1)Determining a threshold for the data.

2)Any data points below the lower threshold are set to the lower threshold, and data points above the upper threshold are set to the upper threshold.

Here are the histograms for the capped features in the dataset. Let's compare them with the original histograms and provide observations:

Distribution Shape: Post-capping, the histograms have retained their overall shape, but extreme values have been limited, making the distributions appear smoother.

Reduction in Outliers: The capping method has effectively reduced the influence of outliers, as evidenced by the smoother tails in the capped histograms compared to the original ones.

Consistency in Distributions:

Features like annualised consumption, power, and efficiency still display a right-skewed distribution. This indicates that the bulk of the data points are clustered towards the lower end, with fewer high values.

The distribution for Emission Intensity remains relatively bell-shaped, suggesting a consistent emission intensity over time.

Improved Clarity: With the reduction of extreme values, the histograms now provide clearer insights into the central tendencies and spreads of the features. This can be especially helpful when building predictive models, as extreme values can unduly influence model performance.

Preservation of Information: While the capping method has limited extreme values, the primary characteristics of the distributions are preserved, ensuring that the essential information in the data remains intact.

In conclusion, capping is a useful method to limit the influence of extreme values without significantly altering the underlying distribution of the data. This can lead to more robust statistical analyses and improved model performance.

What is Log Transformation?

Log transformation is a mathematical operation applied to each data point in a dataset. Specifically, it involves taking the natural logarithm of each data point.

Reasons for Choosing the Log Transformation:

Normalizing the Distribution: Log transformation can make skewed distributions more symmetric, approximating a normal distribution. This is particularly beneficial for many statistical techniques that assume normality.

Decreasing the Effect of Outliers: By compressing the scale, log transformations can reduce the influence of extreme values (outliers). This can make patterns in the data more interpretable and models more stable.

Multiplicative to Additive: Log transformation can convert multiplicative relationships to additive relationships, which can be easier to model and interpret.

Here are the histograms for the log-transformed features in the dataset. Let's compare them with the capped histograms and provide observations:

Distribution Shape:

After the log transformation, many of the features now show a more bell-shaped or symmetric distribution, indicating a closer approximation to normality compared to the capped data.

Reduction in Skewness:

Features that were previously right-skewed (like annualised consumption, power, and efficiency) now appear more centered, suggesting a significant reduction in skewness.

Clarity:

The histograms are less spread out after the log transformation, providing clearer insights into the central tendencies of the features.

Retained Patterns:

The underlying patterns and structures in the data remain largely intact, even after the log transformation.

Improved Interpretability:

The log transformation compresses the scale, making large differences in the original data appear smaller and small differences appear larger. This can help in visualizing and interpreting patterns in the data, especially when the original scale has a wide range.

In conclusion, the log transformation has made many of the dataset's features appear more normally distributed, which can be beneficial for many statistical analyses and modeling techniques that assume normality. The transformed data also provides clearer insights into the central tendencies and spreads of the features.

---------

A correlation matrix is a table showing correlation coefficients between sets of variables. Each random variable (Xi) in the table is correlated with each of the other values in the table (Xj). This allows you to see which pairs have the highest correlation.

This matrix provides pairwise correlation values for the selected variables. The values range from -1 to 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation.

Highly Correlated Variables:

power.GUESS..GW, annualised.consumption.GUESS..TWh, Estimated..MtCO2e, Coal.only..MtCO2e, and Hash.rate.MH.s are highly positively correlated with each other. This means as one of these variables increases, the others tend to increase as well.

This could indicate that as the estimated power used in bitcoin mining increases, the estimated carbon emissions and the hash rate (a measure of mining computational power) also tend to increase. This suggests a direct link between energy consumption, carbon footprint, and the computational power of the bitcoin network.

Emission Intensities:

Emission.intensity..gCO2e.kWh does not have strong correlations with most of the other variables. This might suggest that the intensity of emissions (emissions per unit of energy) remains relatively constant regardless of the other fluctuations in the network.

Efficiency:

The Estimated.efficiency..J.Th variable is weakly correlated with most of the other variables. This suggests that the efficiency of mining hardware (in terms of energy consumed per transaction) might not be a dominant factor influencing the overall energy consumption and emissions of the bitcoin network. However, it's notable that efficiency hasn't drastically improved or worsened significantly over time.

Hydro vs Coal Emissions:

Hydro.only..MtCO2e has weak correlations with the other variables, indicating that emissions from hydroelectric sources don't play a significant role in the overall carbon footprint of bitcoin mining.

On the other hand, Coal.only..MtCO2e is highly correlated with overall estimated emissions (Estimated..MtCO2e). This suggests that coal-based power sources might be a significant contributor to bitcoin's carbon footprint.

## Spatial Analysis

This plot visualizes the monthly absolute hashrate for different countries using a choropleth map, a type of thematic map where the areas shaded based on the value of a variable. from looking at the map we can say that hashrate isn't evenly distributed globally but is concentrated in specific regions or countries due to various factors like technology infrastructure, regulations or energy costs.The colour gradiant ranges from light blue (reprenting lower values) to darkblue (representing higher values) which is used to represent the monthly absolute hashrate value for each country. Any country without data(Na values) is colored in grey. The countries which are shaded dark blue, it would suggest that region as a whole is a significant contributor to hashrate. The countries with higher hashrates might be inferred to have a more significant investment or infrastructure related to the domain.

As of Jan 2022, the maximum contribution in monthly hash rate was of United States followed by China.

Except for US and China, most countries have had a monthly hash rate contribution of 15% or less.

This answers that mining activity varies with regional variations and with the countries listed, being the top contributors.