Correlation Analysis

Nicholas Bradley

# Correlation Analysis

Time series data is continous and chronologically ordered, so there is likely to be some degree of correlation between the series observations. For example, the temperature in the next hour is not a random event since, in most cases, it has a strong relationship with the current temperature or temperatures that have occurred during the past 24 hours. In many cases, the series of past observations contains predictive information about future events, which can be utilised to forecast the series’ future observations. This analysis covers the following topics:

* Casuality versus correlation
* The autocorrelation and partial autocorrelation functions
* Data visualisation tools for correlation analysis

## Correlation between two variables

A principle goal of correlation analysis is to identify and quantify the relationship between two variables, a relationship which well could vary from having a full dependence or linear relationship between the two, to complete independence. A popular method to measure this correlation is the Pearson correlation coefficient, as it is a simple and intuitive representative of statistical logic beyond most methods that measure correlation, although it may not be the most appropriate for time series data. The values of the correlation coeffiecient segment correlation level into three main groups:

* **Positively correlated:** This is where the coefficient value is greater than 0. This indicates some degree of positive linear relationship between the variables depending on the value of the coefficient. As the correlation coefficient grows closer to 1, the linear relationship between the two variables grows stronger; 1 indicates perfect linear dependency.
* **Negatively correlated:** This is where the coefficient value is lower than 0 and it indicates an inverse linear relationship. It basically means that when the value of one variable increases, the value of the other variable decreases. -1 represents a perfect inverse linear dependency between the variables.
* **Not correlated:** This is where the value of the coefficient is equal to 0, which indicates that the two variables are independent.

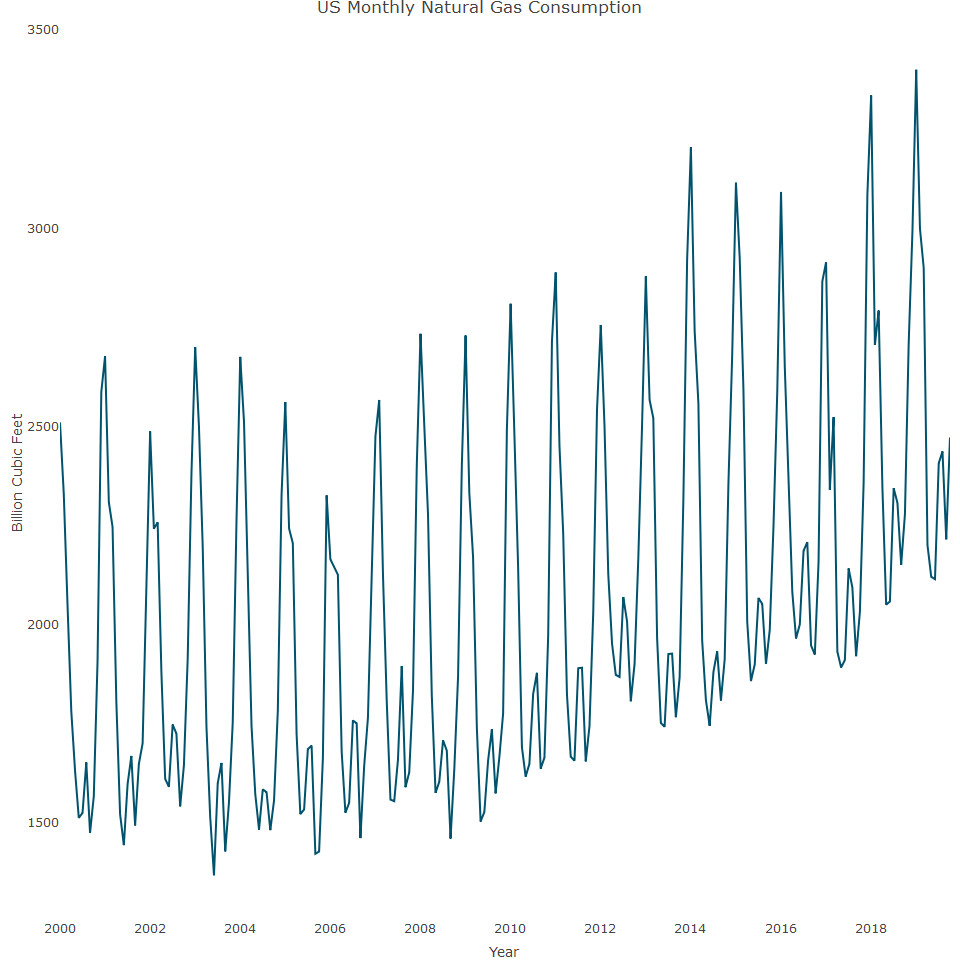
Where two random variables are independent, we can conclude that the correlatin between them is zero. However, if the correlation between two random variables is zero, we cannot conclude by default that the variables are independent as a non-linear relationship may exist. Typically, the correlation between two variables can be considered strong when it is higher than 0.75 and less than -0.75. Measuring and analysing correlation between two variables in the context of time series can be categorised into two categories:

* Analyzing the correlation between a series and its lags, as some of the past lags may contain predictive information, which can be utilised to forecast uture events of the series. One of the most popular methods for measuring the level of correlation between a series and its lags is the autocorrelation function.
* Analysing the correlation between two series in order to identify exogenous factors or predictors, which can explain the variation of the series over time (for example, the effect of weather patterns such as rainfall or temperature on taxi rides in New York City). In this case, correlation measurement is typically done with cross-correlation function

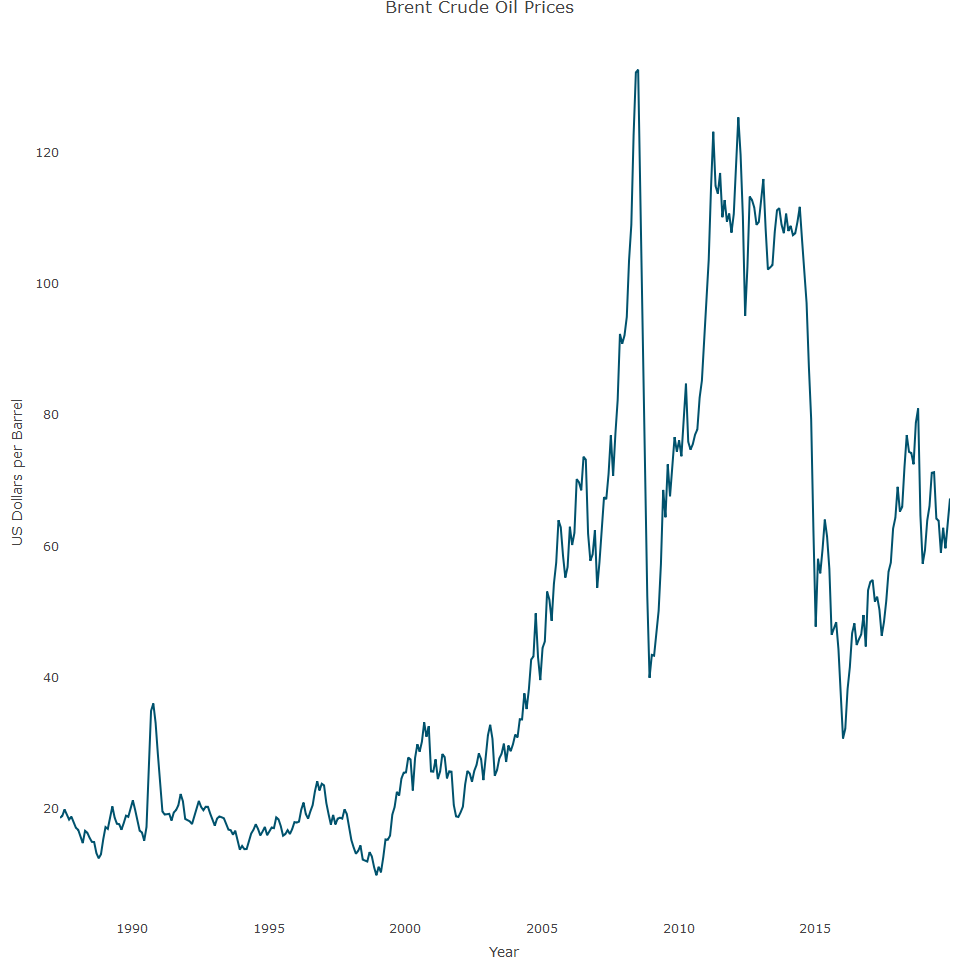
For time series analysis, the first method (also known as lags analysis) is an integrated part of the analysis, while the second method (also known as causality analysis) is less common. This is mainly due to the cost associated with each method. While in lags analysis, we extract the required data (the series lags) from the series itself, causality anlysis requires additional effort such as idenifying and extracting external variables, which may not always be available. We can start with lags analysis methods and then generalise them to the causality analysis approaches.

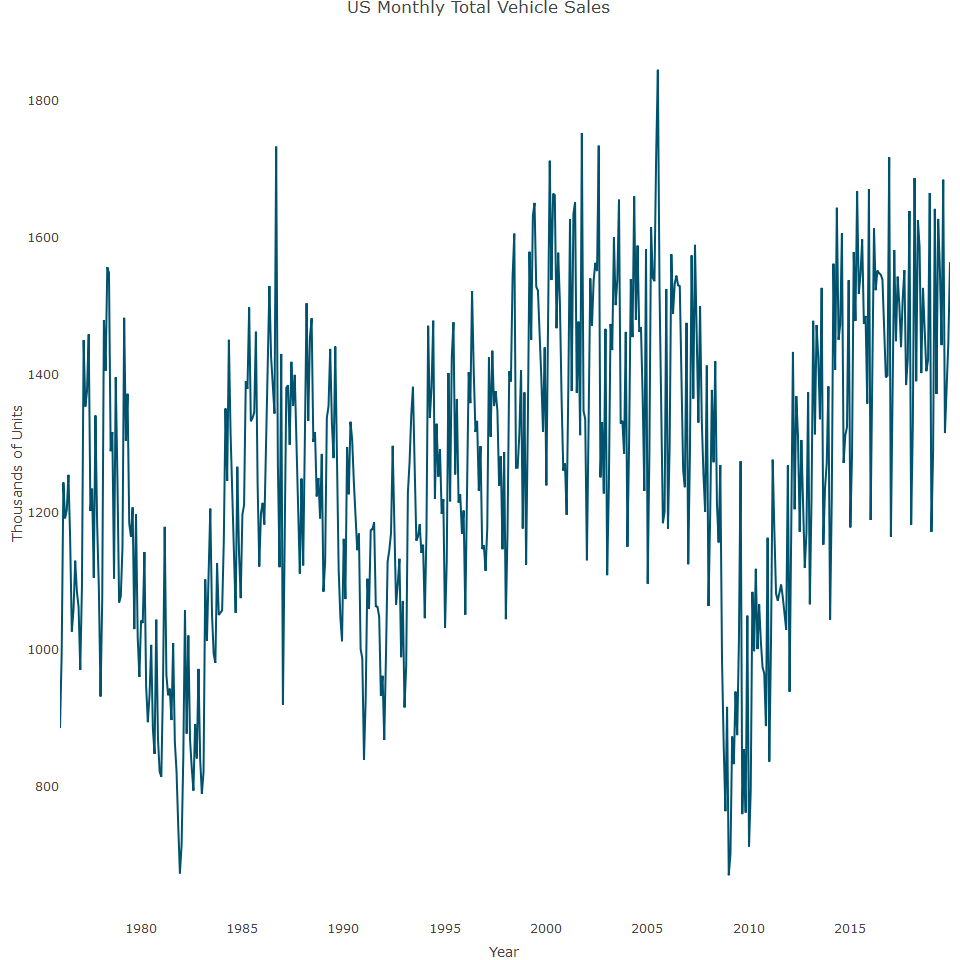
## Lags analysis

The goal of this analysis is to identify and quantify the relationship between a series and its lags and the relationship is generally measured by calculating the correlation between the two and with the use of data visualisation tools. The level of correlation between a series and its lags is derived from the series characteristics. For instance, one should expect the series to have strong correlation with its seasonal lags (for instance, lags 12, 24, 36 when the series frequency is monthly) when the series has strong seasonal patterns. This should make sense as the series direction is impacted by the seasonal pattern. Another example is the price of a stock over time, which in this case, should be correlated with the most recent lags. In the following examples, we will use the US gas, Euro\_Brent and US V Sales series to demonstrate the correlation pattern that is associated with each type.



As evident, the main characteristic of the US gas series is the strong seasonal pattern.





## The autocorrelation function

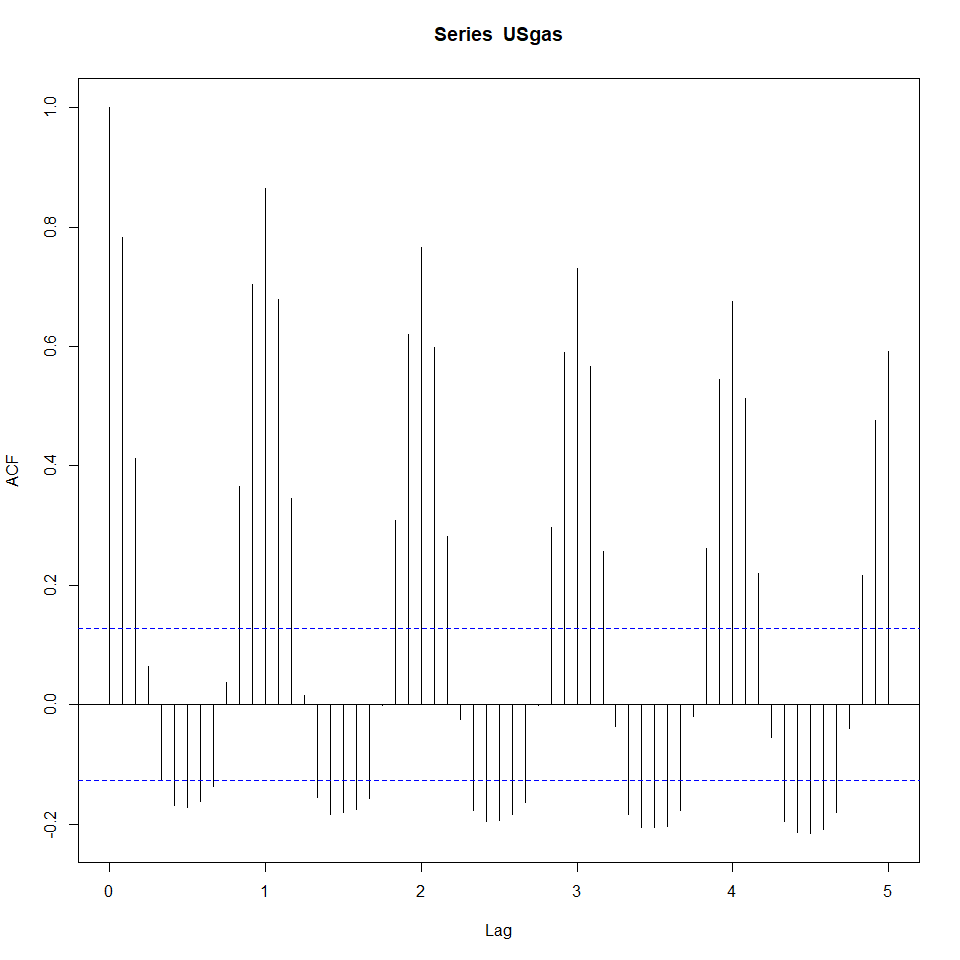
The autocorrelation function (ACF), is the main method in time series analysis for quantifying the level of correlation between a series and its lags. This method is fairly similar (mathematically and logically) to the Pearson correlation coefficient. The Pearson correlation coefficient, often denoted as , measures the strength and direction of the linear relationship between two variables and . It is defined as:

where:

* is the number of pairs of scores
* and are the variables
* is the sum of the product of and
* is the sum of scores
* is the sum of scores
* is the sum of squared scores
* is the sum of squared scores

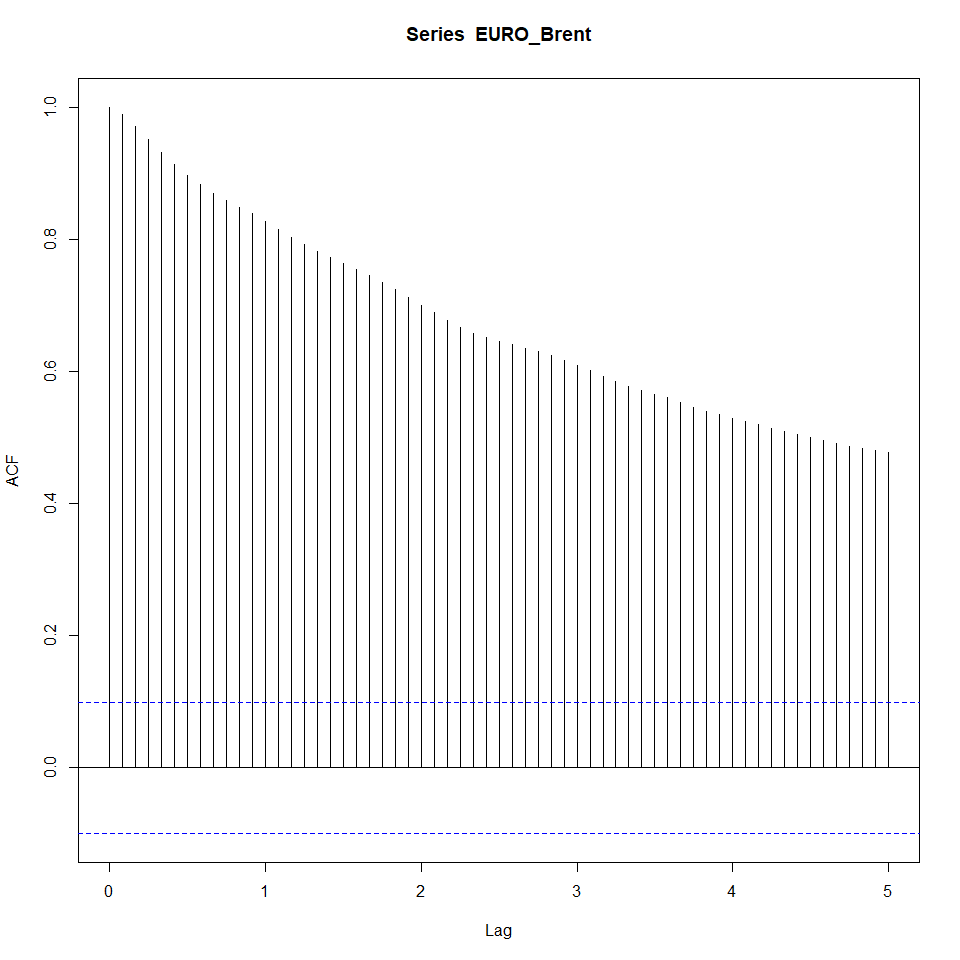
The Pearson correlation coefficient ranges from -1 to 1, where: - indicates a perfect positive linear relationship - indicates a perfect negative linear relationship - indicates no linear relationship

Let’s plot the correlation of the US gas series and its first 60 lags.



Each bar in the ACF plot represents the level of correlation between the series and its lags in chronological order. The x-axis is a bit misleading as the units represent the seasonal lags (for example, lags 1 and 2 represent the 12 and 24 lags). The blue dotted lines indicate whether the level of correlation between the series and each lag is significant or not. By testing the null hypothesis that the correlation of the lag with the series is equal to zero, we can reject it whenever the level of correlation is either above or below the upper and lower dotted lines with a significance of 5%. Otherwise, whenever the correlation is between the upper and lower dotted lines, we fail to reject the null hypothesis and can therefore ignore those lags (or assume that there is no significant correlation between the two)

The above ACF plot indicates that the series has a strong positive correlation with the seasonal lags (which decay over time) along with negative correlation for the mid-seasonal lags (for example, lags, 6, 18, and 30). This should not come as a surprise as this behaviour is aligned with the strong seasonal pattern of the series. Similarly, we can plot the correlations of the EURO\_Brent and US V Sales series and review how their unique correlation pattern aligns with their characteristics



The above ACF indicates that in the EURO\_Brent series, the correlation of the series with its lags is decaying over time whereas the closer the lag is, chronologically to the series, the stronger the relationship with the series. This type of correlation is also an indication that the series is not stationary and a differencing of the series is required.

### Stationarity:

* A stationary time series has statistical properties such as mean, variance, and autocorrelation that are constant over time.
* Stationarity is a crucial assumption in many time series forecasting models, including ARIMA (AutoRegressive Integrated Moving Average).

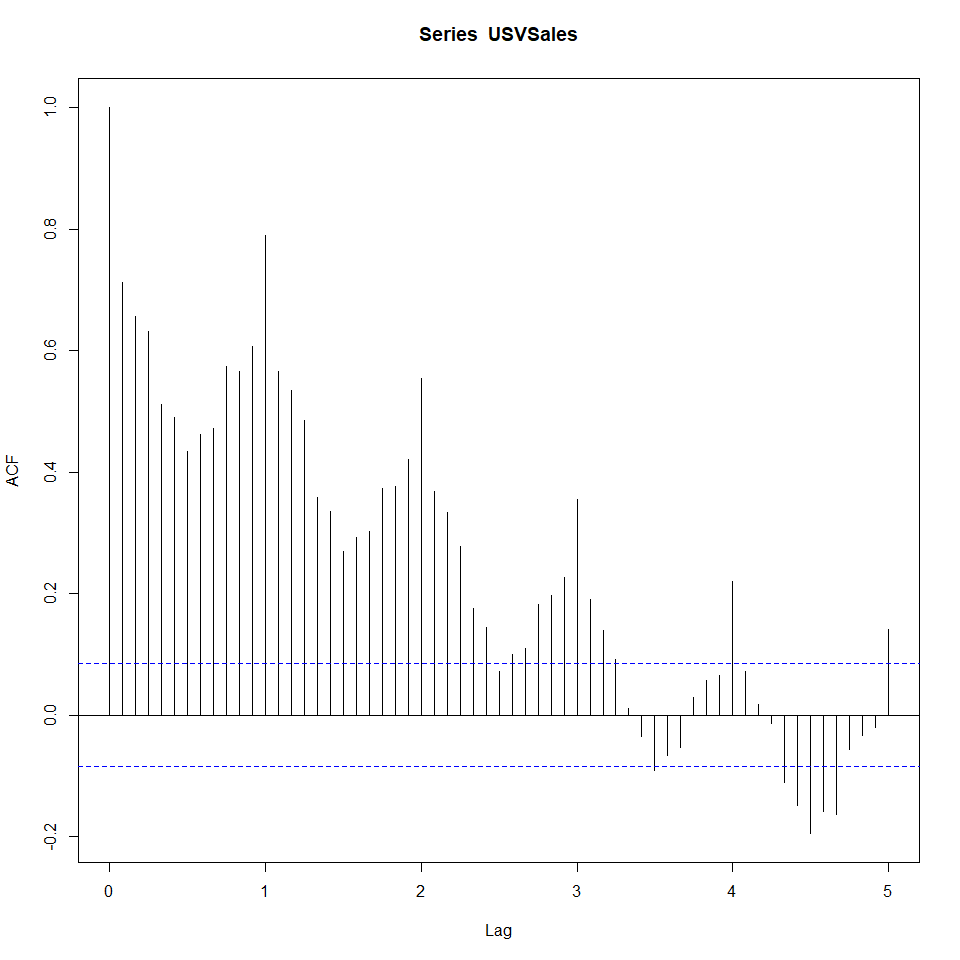
### Non-Stationary Series:

* A non-stationary series has properties that change over time, such as a varying mean or variance.
* Common signs of non-stationarity include trends, seasonal effects, or changing variance.

### Differencing:

* Differencing is a method used to transform a non-stationary series into a stationary one.
* It involves subtracting the previous observation from the current observation. This process can be repeated (first-order differencing, second-order differencing, etc.) until the series becomes stationary.

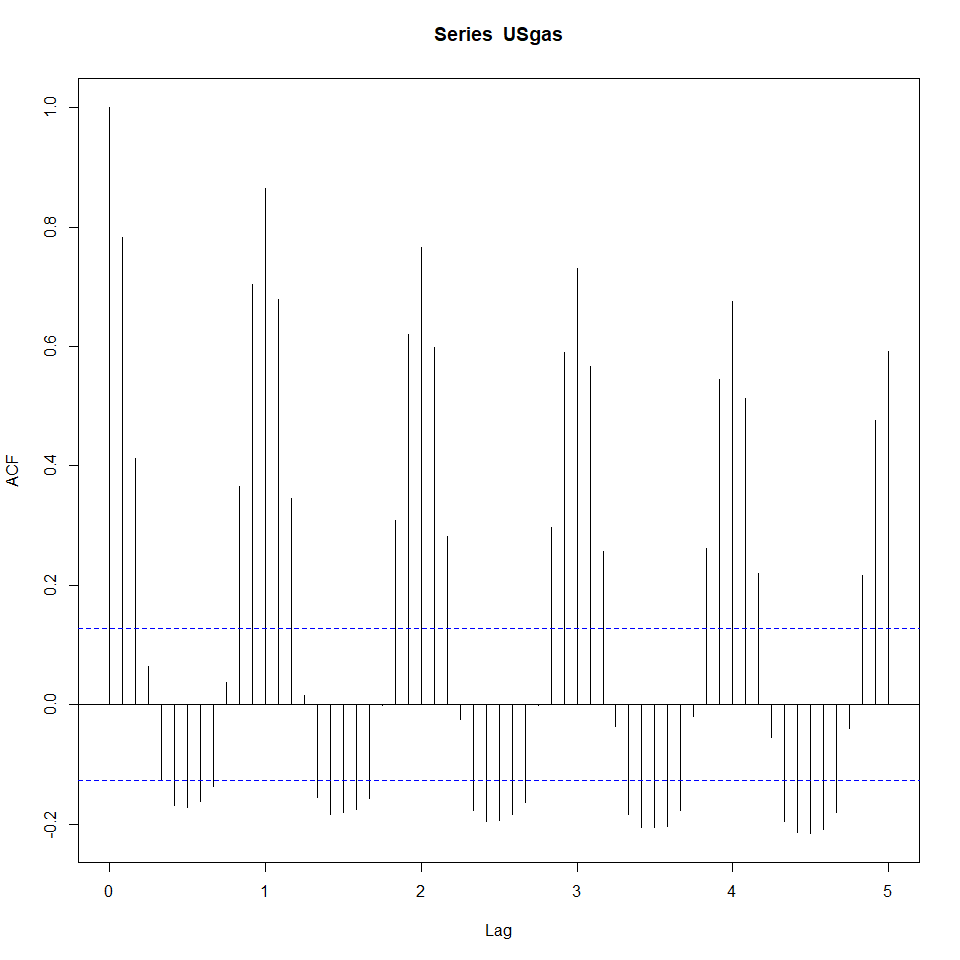
Now let’s generate a plot for USVSales



This correlation has a unique shape, which is a combination of seasonal and cycle patterns of the series. Similarly to USgas, the correlation plot has a cyclic shape as a result of the seasonal pattern of the series. However, the decay rate of USVSales is faster compared to the rate of USgas due to the cycle pattern of the series, which shifts the series direction over time. As a result, the series is mainly correlated with the first seasonal lag. That being said, if we remove the series cycle (or detrend it), we will probably have a similar correlation pattern as USgas.

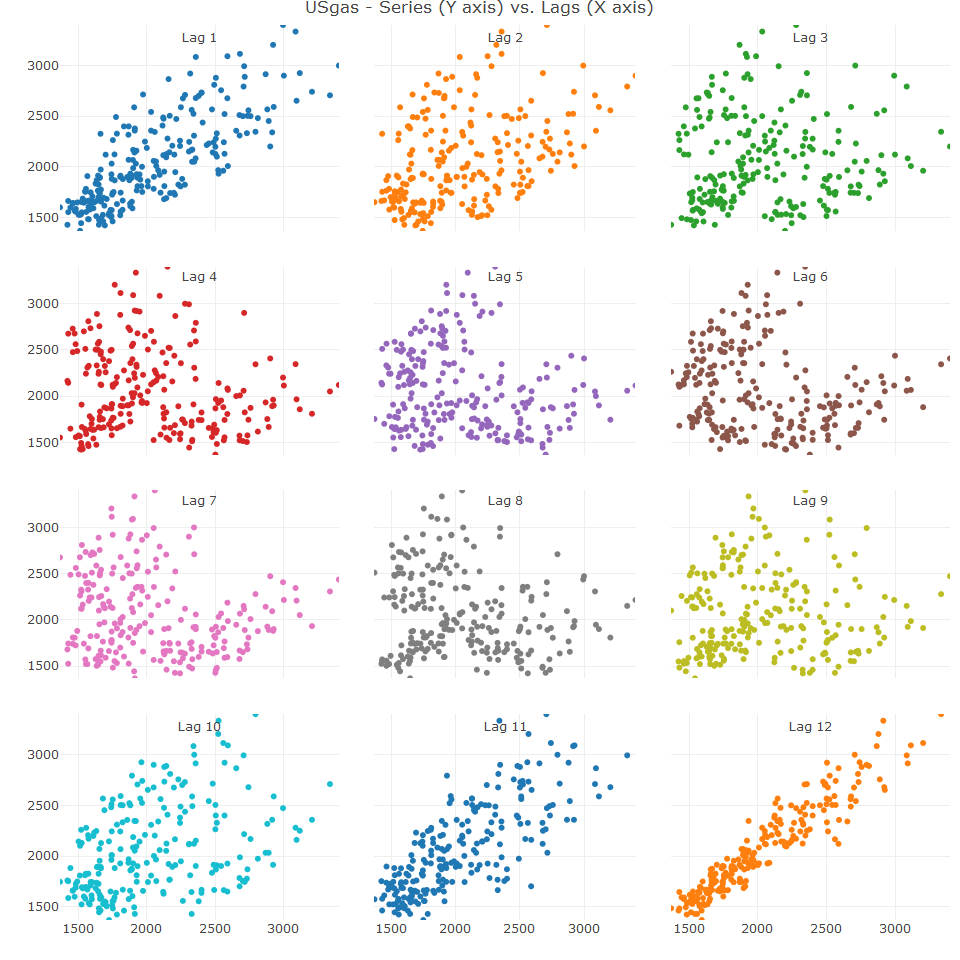
## The partial autocorrelation function

A downside to the autocorrelation function is that it does not remove the effect of lags 1 up to k-1 on the series when calculating the correlation of the series with the k lag. The partial autocorrelation function (PACF), provides a solution by computing the conditional correlation of the series with the k lag given the relationship of the 1,2,…, and k-1 lags with the series. In pther words, the PACF estimates the direct correlation of the series with the k lag after removing the correlation of the k lag with the previous lags. Lets review the PACF output for the first 60 lags of the USgas data



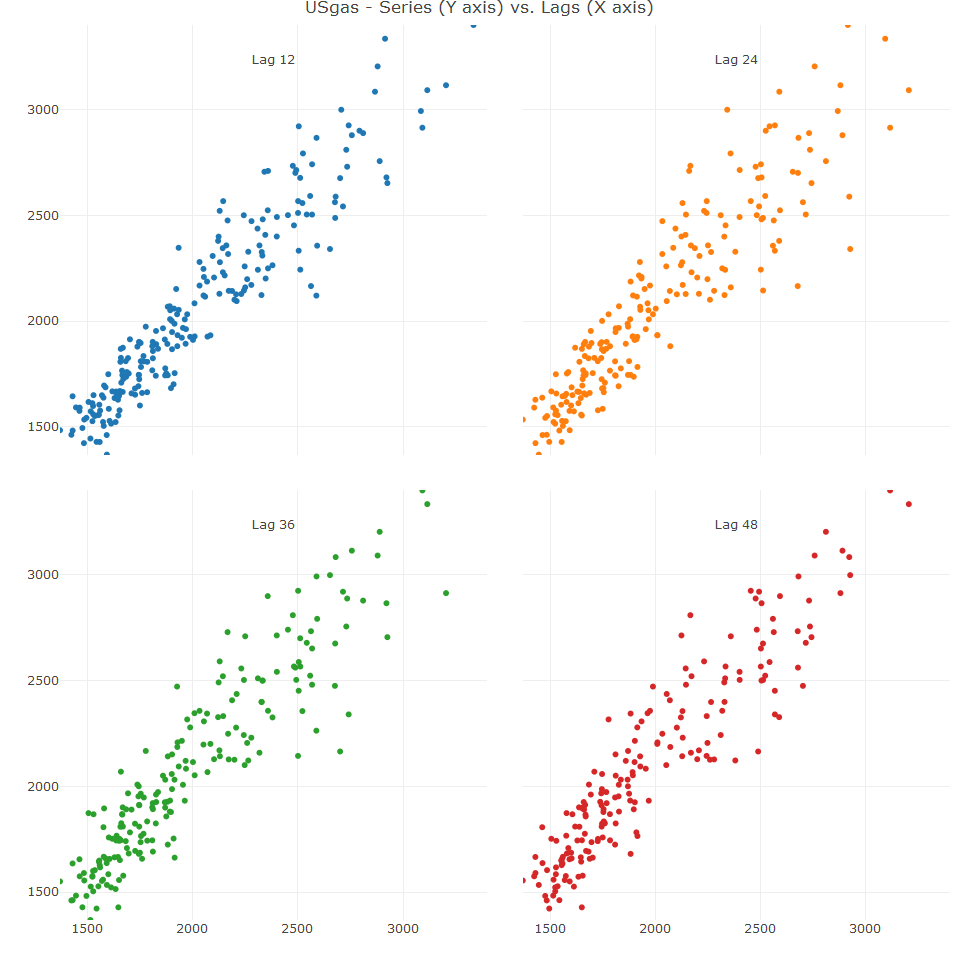
## Lag plots

A lag plot is a simplistic and non-statistical approach for analysing the relationship between the series and its lags. As more points on the lag plot are closer to the 45 degree line, the higher the correlation will be between the series and corresponding lag.



Moving along from the first lag up to the sixth lag, the relationship between the series and its lags becomes less linear. This process start to reverse from the seventh lag as the relationship becomes gradually more linear.

We can plot the most recent seasonal lags (that is 12, 24, 36 and 48) by setting the lags number with the lags argument to the corresponding lags.



## Causality analysis

The use of the series lags to forecast the future value of the series is beneficial wherever the series has stable repeated patterns over time. A good example is the US natural gas consumption as it has a strong seasonal pattern along with consistent trend or growth pattern. However, this method suffers from failing whenever the series changes derive from exogenous factors. In this situation, only using past lags could potentially lead to misleading results as the lags do not necessarily drive the changes in the series. Causality analysis in the context of time series analysis, aims to identify whether a casuality relationship exists between the series we want to forecast and other potential exogenous factors. The use of those external factors as drivers of the forecasting model could potentially provide accurate and robust forecast (as opposed to only using the series past observations).

## Causality versus correlation

Two variables have a causality relationship whenever the change of one variable triggers a direct change in the second variable. Yet, this can sometimes be misleading as correlation itself between two variables does not instantly imply the existence of a causality relationship as the two may have a high dependency on a third variable. For example, in summer, one would expect sales of ice cream and bathing suits to be high during the summer. Yet, there is no causal relationship between the two as both are highly correlated with the same factor - that it is summer.

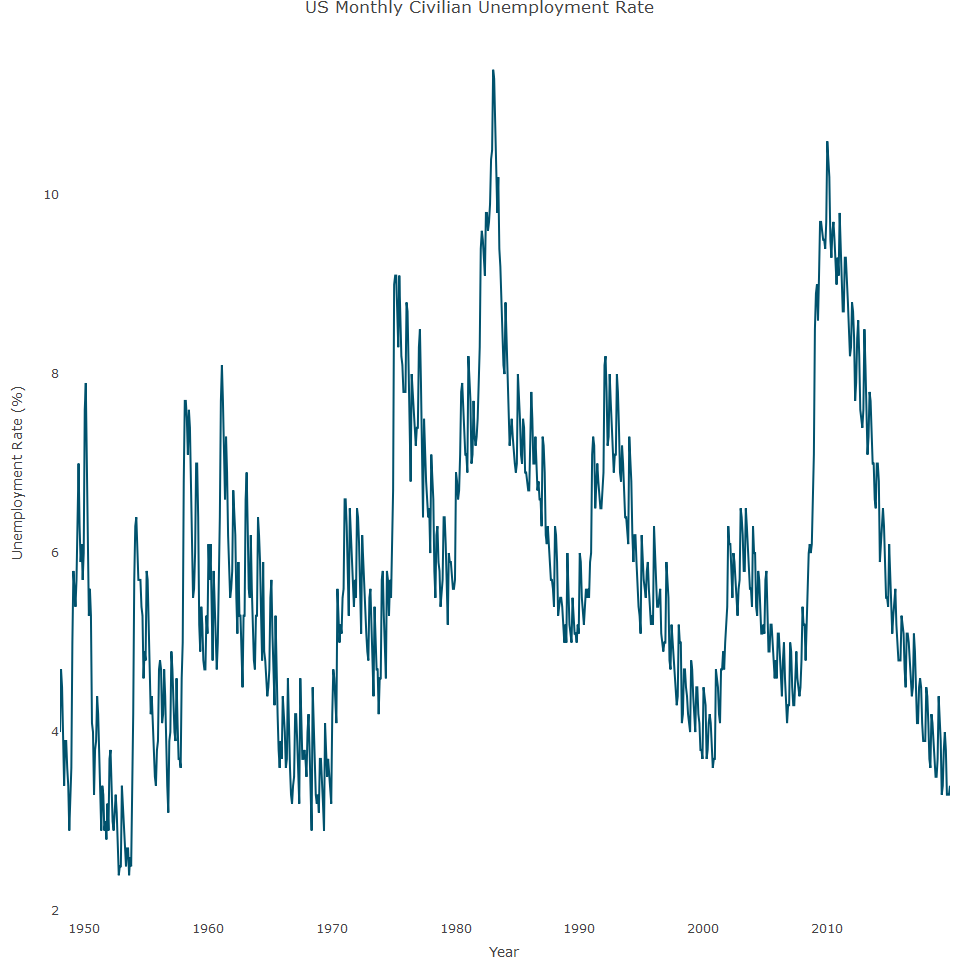
In the context of time series, the causality between two time series can be categorised into the following:

* Direct Causality: Where one series reacts immediately to the change of another time series. For example, wind speed directly impacts on electricity production at a wind turbine.
* In-direct causality: Where the change in one series at time t triggers a change in another series at t + n (where n > 0). This lag effect is common in economic indicators such as Gross Domestic Product (GDP) and the unemployment rate, where a change in one triggers a gradual change in the other over time (for example, a GDP drop this quarter impacts the employment rate in the next quarter)

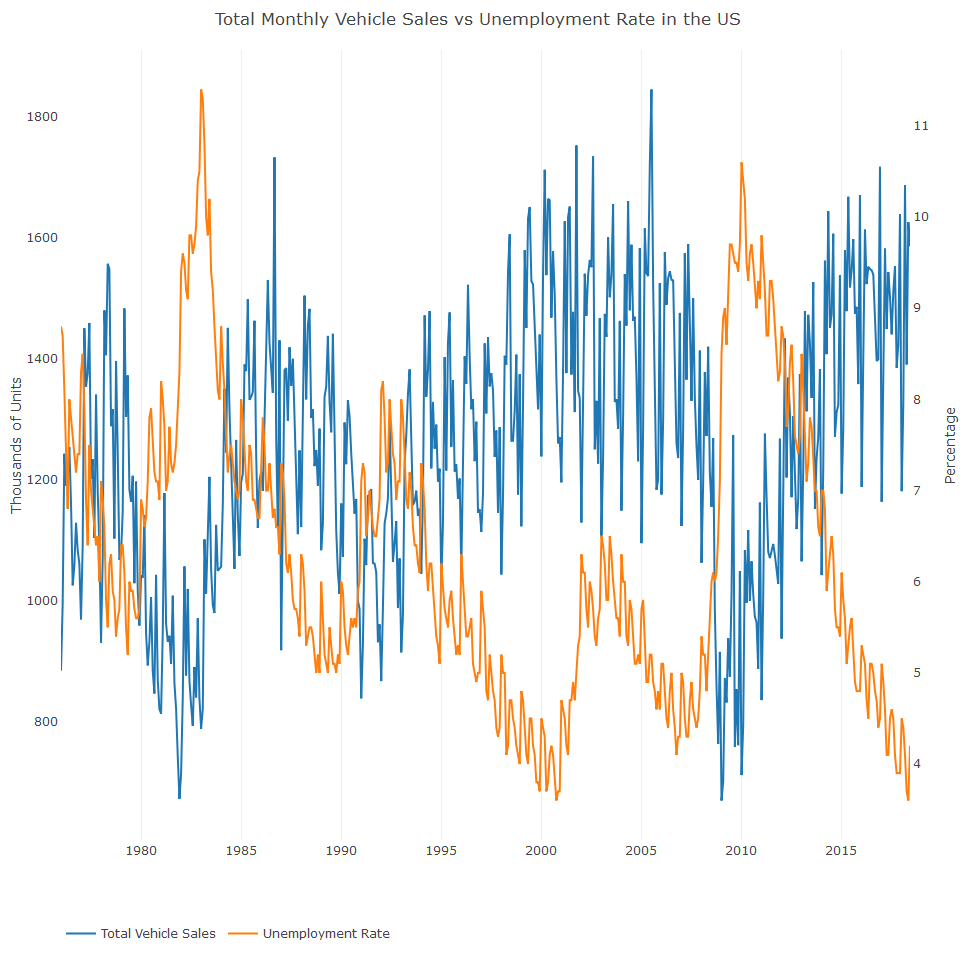
Generally, a series from the first type will have a stronger dependency and a higher level of correlation compared to the second type. Yet, it is harder (or even not practical in some circumstances) to utilize series A as predictor of series b as the future values of series A are unknown and therefore need to be forecas as well (unless the future values are deterministic). This potentially increases the level of uncertainty of the model output due to the fact that the input figures are forecasted and will therefore come with some degree of uncertainty.

## The cross correlation function

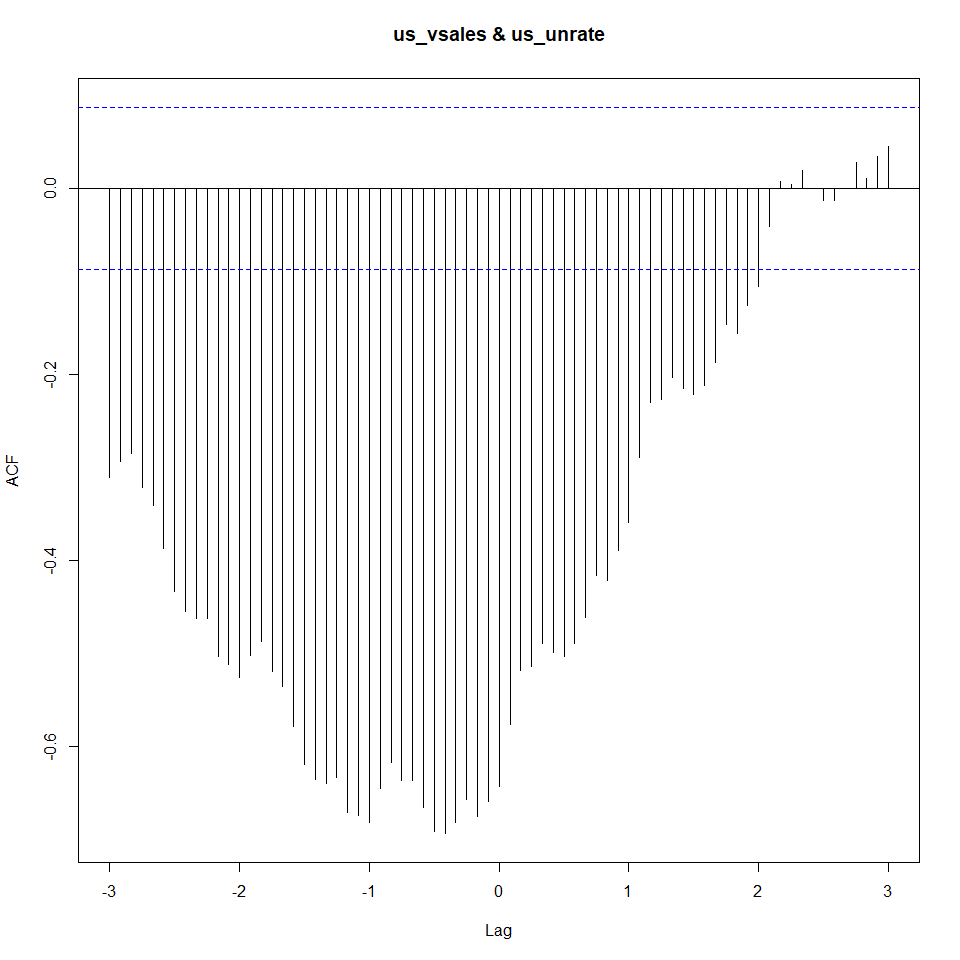
This is the sister function of the ACF and measures level of correlation between wo series and their lags in a similar way.



As seen, the USUnrate begins in the 1950’s, as opposed to the USV Sales series which began in 1976. Let’s align the two series



The plot indicates that the two series move to the opposite direction, so that when vehicle sales increase, unemployment rate decreases and the other way around. In most cases, chanes in vehicle sales series lead to changes in the unemployment rate. In order to explore this assumption further, we can measure the level of correlation between the unemployment rate and the vehicle sales and its lags using the ccf function



Each bar in the CCF plot represents the level of correlation between the main series and the lags of the secondary. Lag 0 represents the direct correlation between the two series, where the negative and positive lags represent the correlation between the unemployment rate and the past and leading lags of the vehicle sales series respectively. The unemployment rate is correlated more with past lags than with leading lags of the vehicle sales. The highest correlation between the two series is found on lag 5 of the vehicle sales and was not far from the seasonal lag as well, that is lag 12. It is difficult and perhaps wrong to conclude from the results that the vehicle sales explictly drive changes in unemployment rate. However, there is some indication of a causality relationship which can be derived from the level of correlation along with common sense, given the size of the vehicle industry in the US and its historical impact on the economy.

Alternatively, it is possible to plot the relationship between US Vehicle sales and the lags of the unemployment rate with the ccf\_plot function.



The advantage of the ccf\_plot function over the ccf function is that the first automatically aligns the series according to their chronological order, whereas the ccf function does not have this automatic functionality and therefore requires preprocessing.